Nouveau design de sondes pour biopuces ADN fonctionnelles et caractérisation des capacités de biodégradation des communautés bactériennes de sols pollués par des hydrocarbures

Sébastien Terrat

UNIVERSITE BLAISE PASCAL **UNIVERSITE D'AUVERGNE**
N° D.U. 2061 ANNEE 2010

**ECOLE DOCTORALE DES SCIENCES
DE LA VIE ET DE LA SANTE
N° d'ordre 528**

THESE

Présentée à l'Université Blaise Pascal
Pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

(Spécialité : GENOMIQUE ET ECOLOGIE MICROBIENNE)

Présentée et soutenue publiquement le
Le 15 Octobre 2010

Sébastien TERRAT

NOUVEAU DESIGN DE SONDES POUR BIOPUCES ADN FONCTIONNELLES ET CARACTERISATION DES CAPACITES DE BIODEGRADATION DES COMMUNAUTES BACTERIENNES DE SOLS POLLUES PAR DES HYDROCARBURES

JURY

Présidente :
Arlette DARFEUILLE-MICHAUD (Pr., JE 2526, Université d'Auvergne, Clermont-Ferrand)

Rapporteurs :
Pascale BAUDA (Pr., UMR CNRS 7146, Université Paul Verlaine, Metz)
Guy PERRIERE (DR CNRS, UMR CNRS 5558, Université Claude Bernard Lyon I, Villeurbanne)

Examinateurs :
Robert DURAN (Pr., UMR IPREM 5254, Université de Pau et des Pays de l'Adour, Pau)
Pierre PEYRET (Pr., UMR CNRS 6023, Université d'Auvergne, Clermont-Ferrand)

Directeur :
Eric PEYRETAILLADE (Dr., UMR CNRS 6023, Université d'Auvergne, Clermont-Ferrand)

Laboratoire Microorganismes : Génome et Environnement – UMR CNRS 6023
Equipe Génomique Intégrée des Interaction Microbiennes
Université Blaise Pascal, Clermont-Ferrand

REMERCIEMENTS

Je souhaite exprimer mes remerciements à Christian Amblard, directeur du laboratoire Microorganismes, Génome et Environnement (UMR CNRS 6023), pour m'avoir accueilli au sein de son laboratoire et permis de réaliser cette thèse dans de bonnes conditions.

Je tiens également à remercier le professeur Pierre Peyret, pour m'avoir accueilli dans son équipe. Merci pour sa grande disponiblité et pour nos nombreuses réunions afin d'orienter au mieux les travaux de cette thèse. Je lui dis merci également pour m'avoir fait profiter de ses connaissances scientifiques qui m'ont beaucoup aidé, et m'ont donné le goût de la recherche, et ce depuis le début de mes études à l'IUT.

J'adresse un très grand merci à mon directeur de thèse, Eric Peyretaillade, tout d'abord pour m'avoir encadré scientifiquement et humainement tout au long de cette thèse. Je veux également le remercier pour toute l'aide qu'il m'a apporté sur mes travaux, pour sa patience avec un étudiant parfois borné, et surtout pour ses encouragements durant ces quatre années au sein de l'équipe. Je le remercie également pour son soutien durant les moments difficiles de cette thèse, et particulièrement durant mes baisses de moral. Je tenais aussi à le remercier pour son omelette aux cèpes, formidable découverte gustative que j'ai eu le plaisir de manger chez lui.

Mes remerciements vont aussi à Olivier Gonçalves pour sa bonne humeur, son humour (parfois grinçant), sa gentillesse et son soutien tout au long de cette thèse, mais aussi durant mon stage de Master Recherche dans cette équipe.

Je tiens à exprimer mes remerciements les plus sincères à Guy Perrière et Pascale Bauda, pour m'avoir fait l'honneur d'accepter de juger ce travail, ainsi qu'à Arlette Darfeuille-Michaud et Robert Duran pour avoir accepté d'examiner ce travail et de faire partie de ce jury de thèse.

Je remercie Julien Troquet, directeur de Biobasic Environnement, et Cédric Vachelard pour le matériel biologique fourni et les données associées dans le cadre d'un projet commun.

Merci aussi à Delphine Boucher et à Corinne Petit pour leurs connaissances qu'elles m'ont fait partager durant ces années. Merci à Delphine pour l'aide dans la rédaction de ce mémoire. Merci à Corinne pour son humour fracassant pendant ces années de thèse.

Merci à tous les membres de l'équipe GIIM pour m'avoir supporté pendant ces quelques années, et surtout pour avoir accepté mon humour parfois déroutant. Merci à eux pour m'avoir donné un environnement de travail agréable grâce à leur présence, leur bonne humeur et leur soutien. Merci également aux étudiants qui sont passés quelques mois dans l'équipe pour laur aide et leur gentillesse: Nicolas Parisot, Xavier Brotel, Yann Keriou, Mélanie Mitchell, Amandine Olivier, Mathieu Roudel, Maxime Ossedat, et j'en oublie sûrement.

Un merci particulier à Anne Moné et à Brigitte Chebance pour leur gentillesse, leur soutien sans faille, et leur disponibilité tout au long de ces années. Merci à toutes les deux pour avoir été présentes pendant les moments difficiles, et pour leur oreille attentive.

Je souhaite aussi remercier Cécile, Ourdia, Emilie, Eric et Jérémie pour leurs discussions de couloirs, leur gentilesse, leur humour et leur ouverture d'esprit. Merci à eux d'avoir été à la fois mes collègues, mais aussi mes amis. Merci aussi à Aurélie, Marlène, Emeline, Olivier, Séréna ainsi qu'aux autres thésard(e)s, étudiant(e)s ou personnel(le)s du laboratoire LMGE pour leur aide et leur soutien. Un grand merci à Jérôme Brunellière (dit Mr Confidentiel, ou encore Mr Dupont….), pour son amitié, sa bonne humeur, et son aide durant cette thèse. Pour nos compétitions sportives ou vidéoludiques, pour nos discussions interminables, etc… Je n'ai qu'un mot à te dire : Chaussette.

Je n'oublie pas mes amis de Clermont, qui sont là depuis de nombreuses années, et qui m'ont toujours soutenu durant cette thèse. Donc merci à Aurélie, Benoît, Florent, Fabien, Guillaume, Emilie, Julie, Antoine, Florence, Thomas, Anaïs, Julien, Xavier et Bruno, le noyau dur de Clermont-Ferrand, pour nos soirées, nos pokers, nos barbecues. Merci aussi à Valérie, Abdel, Julien, Emilie, Agnès, Rémi et Christophe, pour avoir été là.

Un immense merci à mes parents et à mes grands-parents pour m'avoir soutenu durant mes études, et d'avoir accepté mes choix. Merci pour leur présence, leur aide, leur compréhension et leur amour. Merci aussi à Nadège, sans qui je ne serai pas arrivé à faire cette thèse, pour son soutien et son amour indéfectible. Je leur dédie cette thèse.

Enfin, je tiens à remercier le laboratoire Bristol Myers Squibb pour l'invention de l'Efferalgan®, sans qui cette thèse aurait été encore plus difficile.

Nouveau design de sondes pour biopuces ADN fonctionnelles et caractérisation des capacités de biodégradation des communautés bactériennes de sols pollués par des hydrocarbures

Résumé :
Les activités humaines sont à l'origine de nombreuses pollutions par des hydrocarbures au niveau des écosystèmes, et plus particulièrement au niveau des sols. Afin de préserver la santé humaine et environnementale, il est nécessaire d'éliminer les polluants présents. Dans ce but, les techniques de bioremédiation apparaissent aujourd'hui comme de réelles alternatives aux techniques classiques, invasives et onéreuses. Cependant, l'utilisation optimale de tels procédés nécessite une meilleure connaissance des capacités métaboliques des communautés microbiennes impliquées dans la biodégradation de ces polluants. Dans ce cadre, l'utilisation des biopuces ADN fonctionnelles pour analyser ces écosystèmes semble très appropriée. Cependant, une de ses limitations actuelles est la détermination des sondes, qui ne ciblent que les gènes dont les séquences ont été caractérisées. Pour cela, un outil informatique (Metabolic Design) a été mis au point, afin de déterminer des sondes exploratoires pour biopuces fonctionnelles. L'étude, avec notre biopuce fonctionnelle, des capacités métaboliques de dégradation des HAP de la souche *Sphingomonas paucimobilis* sp. EPA505 a permis de mettre en évidence la sensibilité et la spécificité des sondes développées, ainsi que leur aspect exploratoire. Puis, nous nous sommes attachés à caractériser les capacités métaboliques des communautés bactériennes d'un sol pollué principalement par des HAP, sans *a priori* sur les séquences ou les organismes présents, montrant l'efficacité de notre approche.
Mots-clés : biopuce fonctionnelle, hydrocarbures aromatiques polycycliques, design de sondes, sondes exploratoires.

New probe design for functional DNA microarrays and characterization of biodegradation capacities of bacterial communities from hydrocarbons contaminated soils

Abstract:
Soil ecosystems are sensitive to damage from pollutions, and there is an increasing need to develop better methods for removing pollutants from soils. The removal of pollutants, such as polycyclic aromatic hydrocarbons, by bioremediation, is a less invasive and expensive process than classical decontamination. However, use and optimization of bioremediation treatments require knowledge on metabolic capacites of microbial communities involved in the biodegradation of such pollutants. To assess their huge metabolic potentialities, we need high throughput tools, such as functional microarrays, that allow the simultaneous analysis of thousands of genes. However, most classical functional microarrays use specific probes that monitor only known sequences and so, fail to cover the full microbial gene diversity present in complex environments. We have thus developed a program, named Metabolic Design, to design efficient explorative probes for functional microarrays. Then, we successfully validated our new functional microarray studying metabolic capacities of *Sphnigomonas paucimobilis* sp. EPA505 able to degrade polycyclic aromatic hydrocarbons. Finally, we assessed metabolic capacities of microbial communities in soil, contaminated with aromatic hydrocarbons. Results show that our probe design (sensitivity and explorative quality) can be used to study a complex environment efficiently.
Keywords : functional microarray, polycyclic aromatic hydrocarbon, probe design, explorative probe.

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ANNEXES

LISTE DES ABREVIATIONS

%GC	Pourcentage en Guanine et Cytosine
16S	16 Svedberg
aaUTP	Amino-Allyl-UTP
ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
ADNr	ADN ribosomique
ARN	Acide ribonucléique
ARNa	ARN antisens
ARNm	Acide ribonucléique messager
ARNr	Acide ribonucléique ribosomique
ARNt	Acide ribonucléique de transfert
ATP	Adénosine Triphosphate
BLAST	Basic Local Alignment Search Tool
BTEX	Benzène, Toluène, Ethyl-benzène et Xylènes
CDS	Coding DNA Sequence (ou Coding Sequence)
CLHP	Chromatographie Liquide à Haute Pression
CTAB	Bromure d'hexadécyl-triméthylammonium
C_v	Coefficient de variation
DMSO	Diméthyl Sulfoxide
DNase	Désoxyribonucléase
dNTP	Désoxynucléotide tri-phosphate
DTT	Dithiothréitol
EBI	European Bioinformatics Institute
EDTA	Ethylène Diamine Tétra Acétique
EMBL	European Molecular Biology Laboratory
HAP	Hydrocarbure Aromatique Polycyclique
HTML	Hypertext Markup Language
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
IUPAC	International Union of Pure and Applied Chemistry
Kpb	Kilo Paire de Bases
LB	Luria Bertani
ORF	Open Reading Frame (ou cadre de lecture ouvert)
OGM	Organisme Génétiquement Modifié
OTU	Operational Taxonomic Unit
PCR	Polymerase Chain Reaction (ou Réaction en Chaîne de Polymérisation)
POP	Polluants Organiques Persistants
Rpm	Rotation par minute
SDS	Dodécylsulfate de sodium
SNR	Signal to Noise Ratio (ou signal sur bruit)
SQL	Structured Query Language
TBE	Tris-Borate, EDTA
TE	Tris base, EDTA
Tm	Mlting Temperture (ou température de fusion)
TrEMBL	Translated EMBL
U	Unité
UTP	Uridine Triphosphate
UTR	Untranslated Region (ou Région non Transcrite)
UV	Ultra Violet
X-Gal	5-bromo-4-chloro-3-indolyl- -D-galactopyranoside

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Depuis le début de l'ère industrielle, les activités humaines ont engendré la production et/ou l'extraction de composés chimiques, dont la plupart présentent une toxicité reconnue. C'est notamment le cas des métaux lourds, des solvants halogénés, des hydrocarbures ou des pesticides, pour ne citer que les plus couramment rencontrés. Malheureusement, ces activités anthropiques sont également responsables de la dissémination de ces composés dans l'environnement, augmentant les nuisances, et donc les risques pour les individus, mais aussi pour les écosystèmes (Scullion, 2006). Le ministère français chargé de l'Écologie et du Développement durable estime ainsi que 30 000 décès prématurés par an et 7 à 20 % des cancers seraient potentiellement liés à des facteurs environnementaux, comme les pollutions de sources diffuses (transports, utilisation de pesticides) et localisées (incinérateurs, décharges, sites industriels) (source : www.developpement-durable.gouv.fr, données datant de 2009). Cette accumulation de polluants est fréquemment rencontrée dans les sols. En effet, ces écosystèmes sont aujourd'hui les premiers touchés par les activités anthropiques (Scullion, 2006).

Les sols sont également les supports trophiques de la production végétale, les déterminants essentiels de la sécurité alimentaire et de la qualité de l'eau. De plus, en raison de leur position dans l'environnement, les sols jouent un rôle primordial dans les grands cycles biogéochimiques (Eldor A., 2007). Pour cela, et en prenant en compte leur caractère quasiment non renouvelable à l'échelle des générations humaines, les sols constituent un patrimoine dont la gestion durable doit s'imposer dès aujourd'hui. Or, selon la base de données BASOL (données de 2009), on recense 4 186 sites pollués ou potentiellement pollués en France, pour lesquels l'Etat a entrepris une action à titre préventif ou curatif. Dans la majorité des cas, les contaminations sont essentiellement dues à la présence d'hydrocarbures, qu'ils soient aliphatiques ou aromatiques.

Ces hydrocarbures, et notamment les hydrocarbures aromatiques polycycliques, ou HAP, sont une menace majeure pour l'environnement, de par leur caractère récalcitrant au sein des sols (Phale *et al.*, 2007). En effet, les HAP sont classés comme des polluants organiques persistants (ou POP), car ils possèdent un temps de rétention important dans l'environnement, lié à leur faible solubilisation dans les milieux aqueux et leur adsorption aux particules solides. De plus, de par leur caractère cancérigène et mutagène, ces molécules sont devenues une priorité pour les organisations de santé. Ainsi, 16 de ces HAP ont été reconnus comme polluants prioritaires par l'*American Environmental Protection Agency* (ou EPA)

(Vandecasteele, 2005). Certains HAP font également partie des listes de l'OMS (Organisation Mondiale de la Santé) et de la communauté européenne (Seo *et al.*, 2009). Afin de préserver et de restaurer ces écosystèmes et d'éliminer ces polluants, il est donc nécessaire de développer des méthodes fiables et efficaces de réhabilitation. (Andreoni et Gianfreda, 2007; Gan *et al.*, 2009).

C'est avec cet objectif qu'un arsenal de techniques a été mis en œuvre. Les premières ont exploité des procédés mécaniques, physiques et/ou chimiques afin d'éliminer les contaminations. Grâce à ces méthodes, les polluants ont été immobilisés, extraits et/ou détruits pour réduire leurs impacts sur la santé et sur l'environnement (Khan *et al.*, 2004). Cependant, ces techniques peuvent être très invasives pour les écosystèmes traités, car d'une part, elles nécessitent souvent l'excavation des matériaux pollués, ce qui bouleverse les environnements et, d'autre part, elles peuvent entraîner le rejet de composés toxiques. Il existe cependant une alternative à ces techniques, permettant de limiter ces effets néfastes : la remédiation biologique. Cette approche récente utilise les capacités naturelles des plantes (phytoremédiation) et des microorganismes (bioremédiation), à dégrader les polluants, parfois même jusqu'à leur minéralisation complète (Pilon-Smits, 2005; Andreoni et Gianfreda, 2007; Gan *et al.*, 2009). Cependant, même si la phytoremédiation est potentiellement efficace (Pilon-Smits, 2005), elle est généralement limitée aux zones occupées par les racines des plantes, contrairement aux procédés de bioremédiation utilisant les microorganismes, qui sont peu limités dans l'espace.

La bioremédiation s'appuie donc sur les capacités métaboliques des communautés microbiennes capables de dégrader ces polluants. Afin d'optimiser les efficacités de ces techniques de bioremédiation, il est tout d'abord nécessaire de comprendre le mode de fonctionnement de ces communautés vis-à-vis du polluant. En effet, connaître les voies de dégradation des microorganismes mises en jeu dans l'élimination de composés xénobiotiques et leurs régulations reste un élément indispensable pour améliorer les capacités de biodégradation. Cependant, dans la plupart des écosystèmes naturels, ces études se révèlent être de véritables challenges tant les communautés microbiennes sont denses et diversifiées et interagissent entre elles au sein de *consortia* (Schloss et Handelsman, 2006; Dinsdale *et al.*, 2008). Pour l'ensemble de ces raisons, ces populations microbiennes restent encore majoritairement inconnues, et ne peuvent être facilement étudiées par des techniques culturales classiques (Saleh-Lakha *et al.*, 2005).

Les stratégies de caractérisation de ces microorganismes épurateurs se sont donc progressivement orientées sur l'utilisation de techniques de biologie moléculaire, permettant

de s'affranchir de la mise en culture des microorganismes (Galvão *et al.*, 2005). Les premières techniques ont utilisé les potentialités de la réaction en chaîne de polymérisation (PCR), pour amplifier et isoler des gènes d'intérêts jouant le rôle de biomarqueurs (Amann et Ludwig, 2000). Plus récemment, des techniques dites de « haut débit », comme la métagénomique et la métatranscriptomique ont été développées afin d'accéder de manière encore plus exhaustive à l'immense réservoir génétique des populations présentes au niveau des environnements complexes (Stenuit *et al.*, 2008). Ces approches globales ont permis d'améliorer nos connaissances sur les potentialités métaboliques et la diversité phylogénétique des communautés microbiennes présentes au sein des environnements. Cependant, ces techniques ne donnent qu'un aperçu à un moment donné, de l'état de l'écosystème étudié et permettent difficilement de suivre au cours du temps l'évolution des processus biologiques mis en jeu dans les réactions de bioremédiation.

Une autre avancée en biologie moléculaire a été le développement des biopuces ADN (Stenuit *et al.*, 2008). Les biopuces sont basées sur une hybridation des acides nucléiques dite inverse (comparée au Southern blot), et utilisent des sondes immobilisées (des séquences d'ADN simple brin) spécifiques. Actuellement, il est possible de greffer sur des biopuces plusieurs millions de sondes différentes, permettant l'étude d'autant de gènes en une seule manipulation (Dufva, 2009b; Joux *et al.*, 2010). Cet outil moléculaire semble donc tout particulièrement adapté à l'étude, dans un environnement donné, de la structure des communautés microbiennes (biopuces ADN phylogénétiques) et de leurs potentialités métaboliques (biopuces ADN fonctionnelles), mais également pour en assurer le suivi au cours d'un processus biologique. Cependant, une de ses limitations actuelles est la définition des sondes utilisées. En effet, les outils disponibles pour la détermination de sondes ne permettent de cibler que les gènes dont les séquences ont été caractérisées, la fraction inconnue des microorganismes des écosystèmes complexes est donc ignorée (Militon *et al.*, 2007).

Les objectifs de cette thèse ont donc été de développer et de valider une biopuce exploratoire, capable d'appréhender, au sein d'un écosystème pollué, la diversité génique codant les protéines impliquées dans les processus de dégradation des HAP. Pour réaliser ce travail, le développement d'un outil informatique, appelé Metabolic Design, a tout d'abord été mis en œuvre pour permettre : (i) la reconstruction *in silico* de voies métaboliques, utilisant des données publiques et/ou personnelles ; (ii) la définition de sondes spécifiques et de sondes exploratoires. Une première biopuce fonctionnelle a alors été définie pour identifier les gènes dont les produits sont majoritairement impliqués dans les voies métaboliques de dégradation

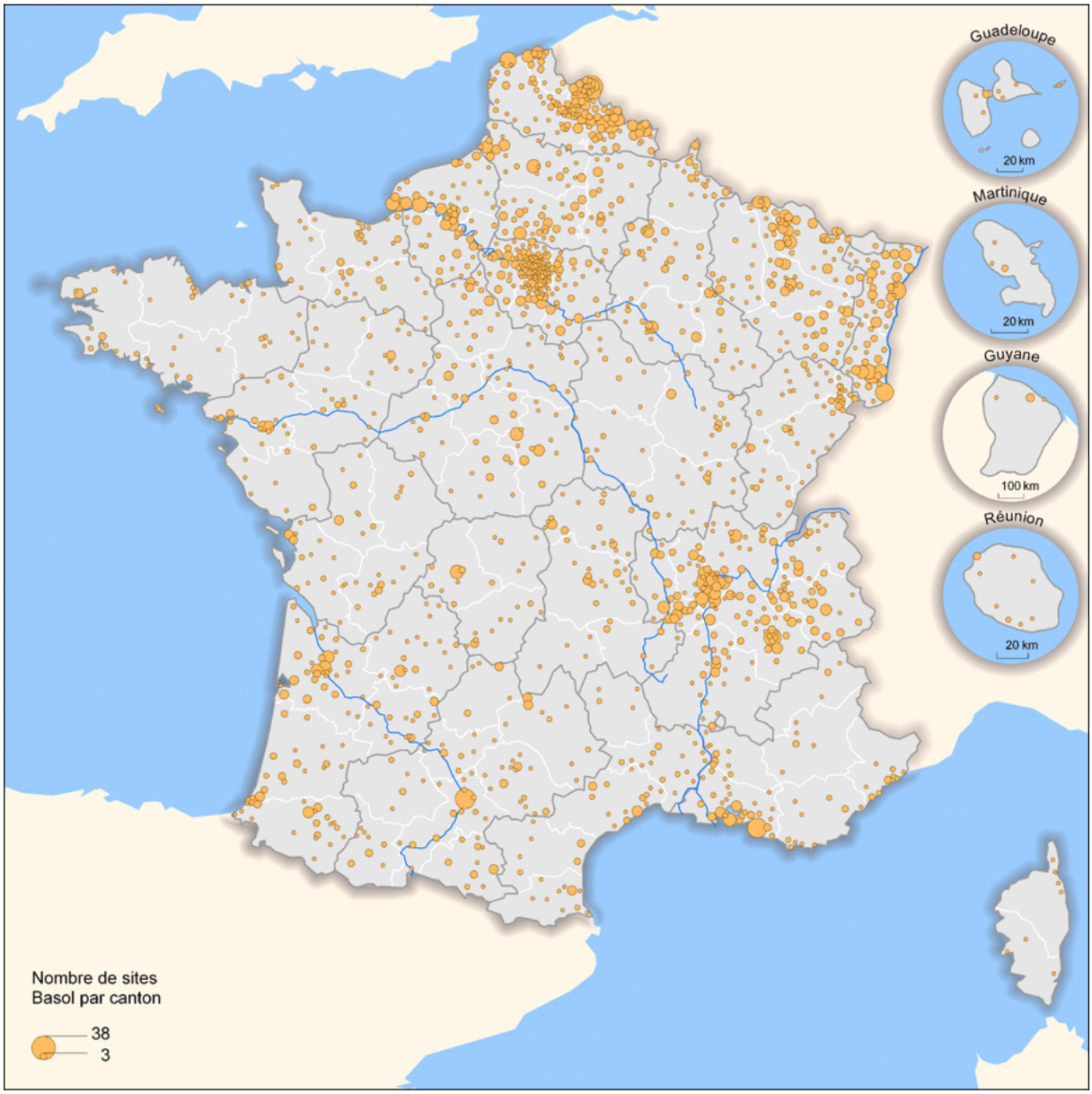
de composés aromatiques, et plus particulièrement celles des HAP. L'utilisation de cette biopuce fonctionnelle pour la caractérisation des potentialités métaboliques de la souche bactérienne *Sphingomonas paucimobilis* sp. EPA505 a permis de valider l'approche. Dans un deuxième temps, cette biopuce exploratoire a été utilisée pour déterminer les capacités métaboliques microbiennes d'un écosystème contaminé par des HAP.

Afin de présenter le contexte de l'étude et les principaux résultats obtenus, le mémoire de cette thèse s'articulera de la manière suivante. Le premier chapitre de l'état de l'art sera consacré à la présentation des sols pollués et des techniques de réhabilitation de ces écosystèmes. Dans le second chapitre, seront décrits les processus biologiques et les potentialités métaboliques de biodégradation des hydrocarbures aromatiques polycycliques par les microorganismes. Dans le chapitre suivant, un état des lieux des approches développées, pour la fouille de données de génomique, et la reconstruction de voies métaboliques *in silico*, sera dressé. Enfin, après une présentation des biopuces ADN, le dernier chapitre permettra de montrer le fort potentiel que représentent les biopuces fonctionnelles en décrivant notamment leur utilisation pour appréhender les diversités métaboliques d'environnement contaminés.

La partie « Résultats » présentera les travaux de recherche effectués afin de répondre à la problématique énoncée. Le premier chapitre sera consacré à la conception et au développement du logiciel Metabolic Design, permettant notamment une définition de sondes exploratoires. Dans ce chapitre seront également décrites les stratégies de détermination de sondes pour biopuces, ciblant des gènes codant des protéines impliquées dans la dégradation de composés aromatiques. Dans le chapitre suivant, l'application de la biopuce ADN pour caractériser les potentialités métaboliques d'une souche bactérienne, sera présentée. Enfin, le dernier chapitre décrira l'évaluation des capacités métaboliques microbiennes présentes au sein d'un écosystème contaminé par des HAP. Ces résultats seront également reliés à une évaluation de la diversité phylogénétique de cet écosystème.

L'intégration et l'interprétation de tous ces résultats seront ensuite regroupés sous la forme d'une discussion générale, permettant de tirer des conclusions générales sur ces travaux ainsi que les perspectives envisagées.

SYNTHESE BIBLIOGRAPHIQUE



<u>Figure 1 :</u> Répartition des sites et sols pollués en France, sur lesquels l'Etat a entrepris des actions de dépollution.

<u>Source :</u> base de données BASOL (données de 2009).
(<u>http://www.stats.environnement.developpement-durable.gouv.fr/acces-thematique/sol/le-sol/les-sites-et-sols-pollues.html</u>)

Chapitre I : Sols et réhabilitation

1. Introduction

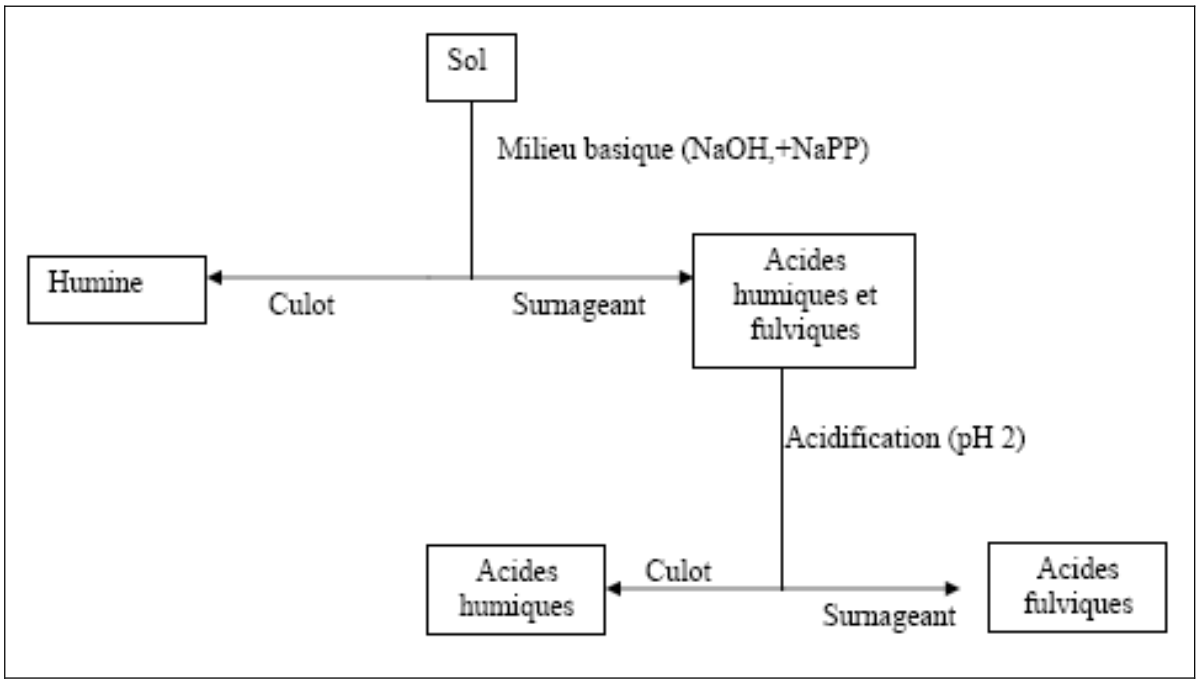
Le sol est la partie supérieure de la croûte terrestre. Cette couche est généralement meuble, superficielle et provient de la transformation de la roche mère (composant la croûte terrestre) par différents processus chimiques et environnementaux (comme l'érosion ou le mouvement des plaques tectoniques). Les sols sont les supports trophiques de la végétation, et à ce titre ont un rôle majeur dans la production de la biomasse. De plus, les sols sont habités par la plus importante biomasse de la Terre : les microorganismes. La surface des sols contiendrait ainsi entre 10^3 à 10^4 kilogrammes de microorganismes par hectare (Fierer *et al.*, 2007b). Cependant, ces sols sont une ressource naturelle non renouvelable à notre échelle de temps, il est donc indispensable de préserver cette richesse pour les générations futures.

Malheureusement, ces écosystèmes sont aujourd'hui les premiers touchés par les activités anthropiques (Scullion, 2006). En particulier, l'industrialisation rapide de l'agriculture, l'expansion des industries chimiques, et la nécessité de produire de l'énergie ont abouti à une dépendance vis-à-vis des composés chimiques d'origine anthropique, et ont entraîné la contamination des sols par un grand nombre de xénobiotiques (Andreoni et Gianfreda, 2007). Selon les bases de données BASOL et BASIAS (données de 2009), on recense ainsi 4 186 sites pollués en France (Figure 1), et 246 000 sites potentiels qui ont abrité, par le passé, une activité pouvant être à l'origine d'une pollution. On retrouve principalement dans ces sites pollués des métaux lourds (cadmium, arsenic, chrome, cuivre, mercure), des solvants halogénés, des hydrocarbures ou des pesticides (Scullion, 2006).

L'intérêt actuel de la recherche porte sur l'étude des polluants organiques et inorganiques car ce sont eux qui se retrouvent en grande partie immobilisés au niveau des sols. Ils engendrent alors de profondes modifications des écosystèmes. C'est pourquoi, il est indispensable de trouver des solutions pour traiter les polluants déjà présents, mais aussi pour diminuer les pollutions futures.

2. Le système sol

Le sol est principalement composé de particules minérales, de matière organique, d'eau, d'air et d'organismes (Eldor A., 2007). Le sol a de nombreuses fonctions. Il est un milieu biologique dans, et sur lequel, se développent des êtres vivants. Ce développement va



<u>Figure 2 :</u> Méthode d'isolement des différentes fractions des substances humiques présentes au sein des sols.

dépendre de la qualité de ce sol, ce qui va définir son niveau de fertilité (quantité de carbone, d'azote, capacité d'échange cationique, etc.…). Il est aussi un acteur déterminant du cycle de l'eau (stockage et régulation), et de la qualité de celle-ci (source de pollution, capacité de rétention des polluants mais aussi biodégradation de ces derniers). Le sol est donc un système hétérogène complexe qui est composé de plusieurs fractions différenciables. Ces fractions se distinguent non seulement par leur origine, mais également par leurs propriétés, entraînant leur intervention à plus ou moins grande échelle dans la sorption des polluants dans les sols (Musy et Soutter, 1993).

2.1. La fraction minérale

La fraction minérale provient de la dégradation physique ou chimique de la lithosphère. Cette dernière peut subir diverses altérations physiques (érosion, désagrégation mécanique, chocs thermiques) et/ou chimiques (hydrolyse, oxydoréduction, dissolution, hydratation/déshydratation, complexation) permettant la formation d'ions, d'oxydes et de minéraux argileux. La fraction minérale est principalement constituée de minéraux primaires (quartz, feldspaths, micas…), et de minéraux secondaires, les oxydes métalliques ou les argiles. Ces minéraux représentent en général 95 à 99 % du sol et peuvent être de tailles très diverses : argile granulométrique (diamètre <2µm) ; limon (diamètre de 2 à 50µm) ; sable (diamètre de 50 à 2000µm).

2.2. La fraction organique

Contrairement à la fraction minérale, la fraction organique n'est pas constituée de particules élémentaires dissociables selon leurs dimensions. Elle est composée à plus de 80 % de matière organique morte, dont les transformations mécaniques et enzymatiques sont réalisées par la microflore, la mésofaune et la macrofaune du sol (Bot et Benites, 2005).

La décomposition de la matière organique entraîne une accumulation de matière non dégradée formant l'humus qui constitue de 35 à 55 % de la fraction organique morte. C'est un composé relativement stable, formé de substances humiques incluant les acides humiques (fraction soluble dans l'eau, sauf à un pH inférieur à 2), les acides fulviques (fraction soluble dans l'eau à n'importe quel pH), les acides hymatomélaniques et les humines (fraction insoluble dans l'eau à n'importe quel pH) (Figure 2) (Bot et Benites, 2005; Eldor A., 2007). Du fait de leurs structures complexes, les substances humiques ne peuvent être facilement utilisées par les microorganismes et restent donc stables durant un temps relativement long bien qu'elles puissent subir une minéralisation secondaire. Ainsi, il est par exemple possible,

par un apport de sources de carbone plus récentes, de faciliter leur dégradation (Fontaine *et al.*, 2007).

2.3. La fraction vivante

La fraction vivante est composée pour une part importante, mais non exclusive, d'une large variété de microorganismes comme les bactéries, les archées, les champignons, les protozoaires, les algues et les virus.

Les bactéries et les archées sont les microorganismes les plus étudiés parmi les groupes microbiens des sols. Certaines estimations indiquent que 10^3 à 10^6 « espèces » procaryotiques pourraient être présentes dans un gramme de sol, ce qui constituerait le groupe d'organismes le plus diversifié (Fierer *et al.*, 2007b). Cette diversité d'espèces très importante met en évidence la grande complexité des communautés procaryotiques présentes au sein de ces écosystèmes, comme le montre de nombreuses études portant sur des sols très divers (Hernandez-Raquet *et al.*, 2006; Fierer *et al.*, 2007a; Kim et Crowley, 2007b; Kim *et al.*, 2008; Youssef et Elshahed, 2008; Kumar et Khanna, 2010; Rastogi *et al.*, 2010). Il est intéressant de noter que bien qu'une grande diversité soit visible au niveau taxonomique de l'espèce, la structure des communautés au niveau taxonomique des *phyla* semble beaucoup stable, et ce malgré les variations de pH, de température ou les propriétés des sols (Youssef et Elshahed, 2008). Cependant, il est très difficile de relier précisément ces différents groupes bactériens aux grandes fonctions biologiques présentes au sein des sols (Fierer *et al.*, 2007a).

Les champignons ont été étudiés pendant des siècles, mais grâce aux dernières avancées en matière de biologie moléculaire, il est maintenant démontré que l'on avait sous-estimé la diversité totale des communautés fongiques des sols (O'Brien *et al.*, 2005; Fierer *et al.*, 2007b). Le rôle des mycètes dans la dégradation de la matière organique est multiple : dégradation de la celllulose mais aussi de la lignine ou des tanins. De récentes études sur les virus (Williamson *et al.*, 2003; Williamson *et al.*, 2005) ont permis de développer des protocoles permettant l'extraction directe, et l'énumération des virus des sols. Les données obtenues ont révélé que les virus sont très abondants dans ces écosystèmes complexes, pouvant même dépasser l'abondance bactérienne (Srinivasiah *et al.*, 2008). Tout comme dans les réseaux aquatiques, les virus participent activement aux transferts de gènes (transduction), mais aussi au maintien des différentes populations bactériennes du sol en décimant celles devenues trop dominantes (théorie du « killing the winner »).

La faune du sol, très importante et très diversifiée, est également impliquée dans la dégradation et la structuration de la matière organique (Eldor A., 2007). Un mètre cube de sol

de prairie pourrait ainsi contenir jusqu'à 260 millions d'individus divisés en quatre classes selon leur taille : la microfaune (< 0,2 mm), constituée majoritairement de Protozoaires, de Nématodes, mais aussi de Rotifères et de Tardigrades ; la mésofaune (0,2 à 4 mm), constituée de Microarthropodes tels que les Acariens et les Collemboles, de Pseudoscorpions, de Protoures, de Diploures, de petits Myriapodes et de vers ; la macrofaune (4 à 80 mm), composée de vers de terre (*Oligochaeta*), d'insectes (comme les fourmis et les termites), de limaces, d'escargots, d'araignées et d'opilions, et la mégafaune (> 10 cm) qui regroupe les mammifères, les reptiles et les amphibiens. (Deprince, 2003).

2.4. Les phases liquides et gazeuses

La « porosité du sol » comprend tout ce qui n'est pas solide et représente en général 30 à 60 % du volume total du sol. La phase liquide est composée d'eau et contient des composés dissous tels que de nombreux ions minéraux provenant de la dissolution des constituants du sol, et des composés organiques et minéraux issus de la décomposition de la matière organique morte. Cette phase liquide peut se retrouver sous forme d'eau de ruissellement (mouvements parallèles à la surface), d'eau de gravité (mouvements verticaux ou obliques) ou d'eau retenue, dans le cas où les forces de capillarité et d'absorption sont supérieures aux forces de gravité (Duchaufour, 2001). La phase gazeuse, quant à elle, est composée d'azote, d'oxygène, de gaz carbonique et parfois de méthane. Les trois phases (solide, liquide et gazeuse) interagissent entre elles et donnent lieu à de nombreuses réactions chimiques. Les différentes interfaces solide-liquide, solide-gaz, liquide-gaz et solide-solide sont les lieux privilégiés de ces réactions.

3. Dynamique et devenir des polluants organiques au niveau des sols

Les sols contiennent une large gamme de composés ayant des propriétés physiques et chimiques différentes (Huang *et al.*, 2003). Cependant, depuis des décennies, ces écosystèmes sont soumis à de fortes pollutions qui sont généralement liées aux activités anthropiques. Or, certaines fractions du sol sont largement impliquées dans l'immobilisation des polluants organiques au niveau des sols, ce phénomène étant appelé géosorption. La géosorption des polluants organiques peut se faire préférentiellement au niveau de la fraction organique des sols, si la composition du sol est riche en carbone organique (supérieur à 0,1 %), ou au niveau de la fraction minérale des sols, si celui-ci est pauvre en carbone organique (Huang *et al.*, 2003; Ehlers et Loibner, 2006).

Le transport, la sorption et la désorption des polluants organiques dans les sols dépend de la nature des polluants considérés, mais aussi de la nature de la morphologie, et des propriétés macro- et microscopiques de la matrice solide (Kan *et al.*, 1994; Ehlers et Loibner, 2006). Les facteurs environnementaux comme la température, le pH et le taux d'humidité peuvent également jouer un rôle majeur, comme les propriétés physico-chimiques des polluants, leur structure moléculaire et la concentration des polluants (Park *et al.*, 2003). Il semble de plus que les polluants organiques tendent à être de moins en moins mobiles avec le temps (van Noort *et al.*, 2003). Ces phénomènes de « vieillissement » affectent ainsi les mouvements et la réactivité des polluants organiques au niveau des sols, et donc leur disponibilité pour les microorganismes (Kan *et al.*, 1994; Huang *et al.*, 2003).

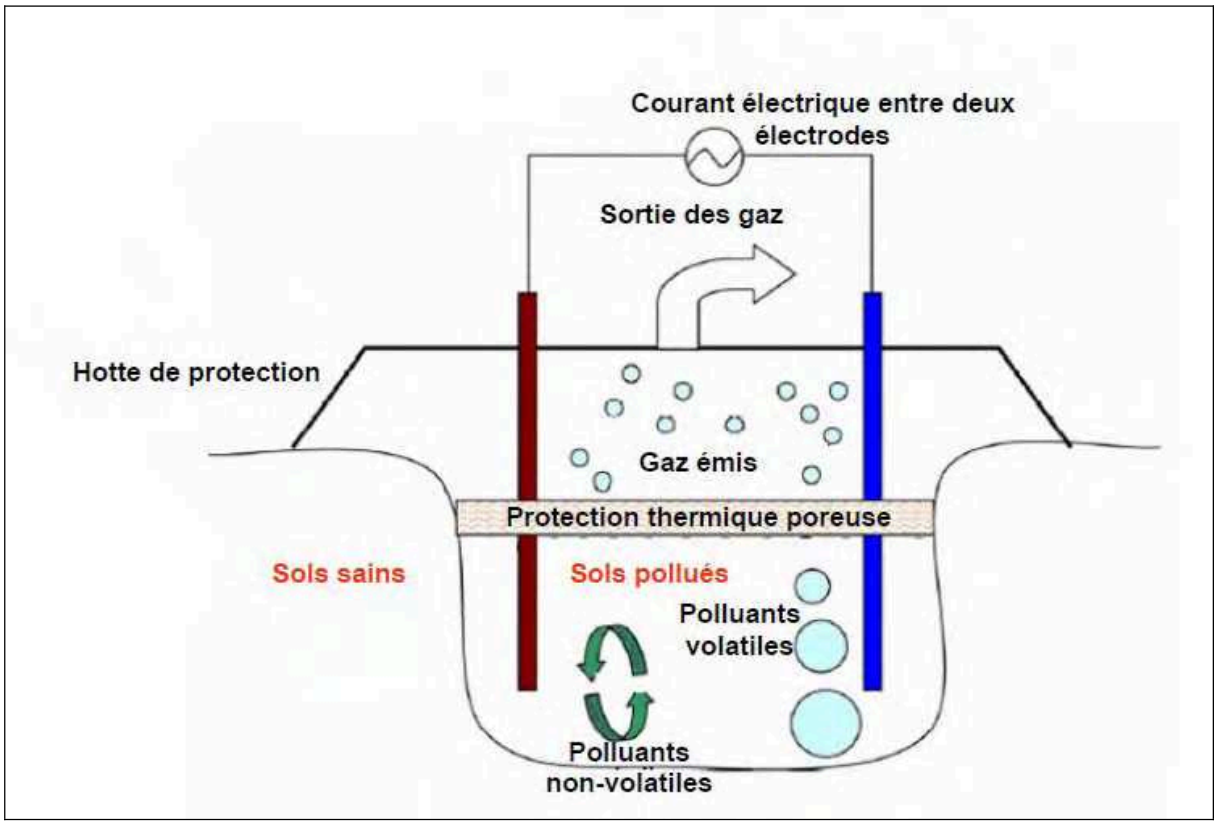
4. Réhabilitation des sols pollués

Les sols pollués ont été traditionnellement excavés pour être stockés dans des décharges, ou alors isolés en utilisant des « barrières » variées (comme des parois imperméables, des films, etc.…) afin d'éviter toute propagation hors site vers des zones peuplées. En termes de gestion des risques, ces approches aidaient à contrôler les pollutions sans les traiter et sans éliminer la source de pollution. De nombreuses techniques de réhabilitation ont alors été développées afin de traiter *in situ* et *ex situ* les sols pour éliminer la pollution (Khan *et al.*, 2004; Scullion, 2006). Les buts de ces traitements peuvent être d'éliminer complètement ou non les polluants, d'extraire et de récupérer les polluants pour un traitement ultérieur, de stabiliser les polluants sous des formes moins mobiles ou toxiques, de séparer les fractions polluées de celles non polluées, et enfin d'empêcher leur diffusion vers d'autres écosystèmes. On peut classer ces techniques de réhabilitation en deux groupes : les techniques non biologiques (qui peuvent être physiques ou chimiques) et les techniques biologiques (qui se basent sur l'utilisation d'organismes vivants, comme les plantes ou les microorganismes).

4.1. Réhabilitation non biologique

4.1.1. Techniques physiques

Une des méthodes physiques couramment utilisée est le traitement thermique des sols *in situ* ou *ex situ* (Khan *et al.*, 2004; Bonnard *et al.*, 2010). L'application de hautes températures (supérieures à 1 000°C) permet la destruction des composés organiques, et le piégeage au sein des particules du sol des composés inorganiques (par modification de leur



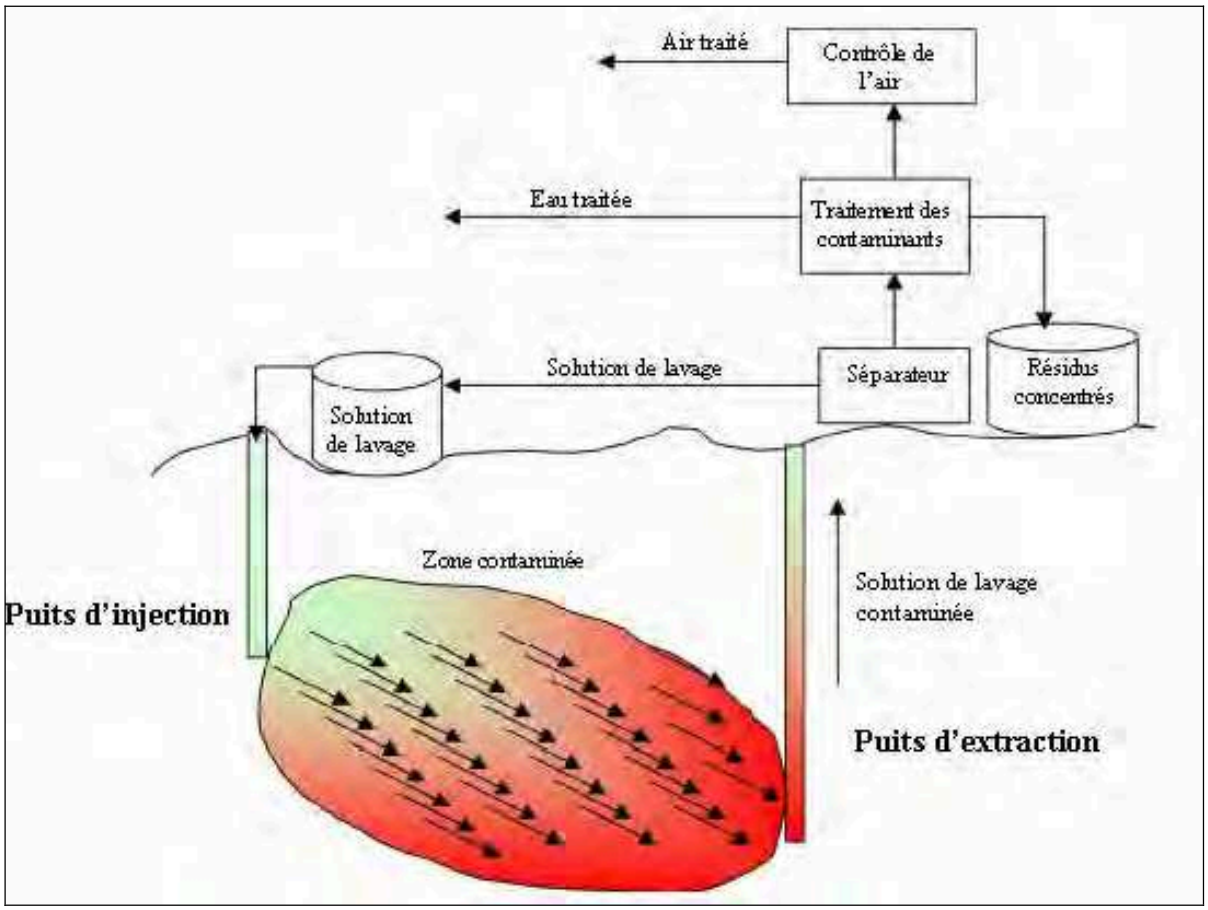
<u>Figure 3 :</u> Principe de la vitrification *in situ* des polluants du sol.
La vitrification est une technique qui permet de transformer, à hautes températures, les sols pollués en pâte de verre inerte. Cette fusion à haute température réalisée sur la zone polluée (entourée par des sols sains) entraîne une dégradation des composés organiques et un piégeage des composés inorganiques (adapté de Khan *et al.*, 2004).

mobilité). Comme les propriétés bio-physico-chimiques de ces terres incinérées ne sont plus identiques à celles de terres non traitées, ces résidus de combustion ne peuvent plus être utilisés comme support de la végétation (Scullion, 2006). Une autre technique thermique plus « douce », appelée désorption thermique, consiste à chauffer les sols pollués excavés jusqu'à la température d'ébullition des polluants, généralement de 100°C à 600°C, afin de les vaporiser (Biache *et al.*, 2008; Bonnard *et al.*, 2010). Ce changement d'état permet de séparer les contaminants de la matrice sol. Ces vapeurs sont ensuite collectées afin d'être traitées. Il est ainsi possible de conserver certaines propriétés essentielles des sols (comme sa porosité, sa perméabilité ou sa texture). Des études récentes démontrent ainsi une forte baisse (de plus de 90 % dans ces études) de la concentration en HAP de ce sol pollué, après traitement.

La « vitrification » est une autre méthode de traitement thermique utilisant des températures comprises ente 1 600 et 2 000°C (Figure 3), qui permet la solidification et/ou la stabilisation des polluants. Cette haute température permet : la volatilisation des composés organiques avec peu de formation de métabolites et l'incorporation dans une « pâte de verre » très solide et résistante au lessivage des polluants inorganiques (Acar et Alshawabkeh, 1993). Une méthode alternative à la vitrification, mais qui permet également la solidification et/ou stabilisation est l'incorporation à l'asphalte. C'est une technique qui consiste à excaver les sols pollués par des hydrocarbures et à les substituer aux granulats lors de l'élaboration de l'asphalte (Khan *et al.*, 2004; Scullion, 2006). Les mélanges inertes et imperméables alors obtenus sont ensuite utilisés, par exemple, pour la construction des infrastructures de transport.

Les réhabilitations des sols par injection de vapeur d'eau ou d'air chaud sont des traitements basés sur la mise en circulation des polluants entre les phases liquide et gazeuse ou leur volatilisation (Halmemies *et al.*, 2003; Scullion, 2006). L'injection de vapeur d'eau, ou d'air à haute température, puis la récupération de ces gaz et leur traitement ultérieur permet une élimination pratiquement complète des polluants ciblés. Ces techniques peuvent être appliquées *in situ*, n'entraînant que peu de perturbations des sites traités (Halmemies *et al.*, 2003). Ce type de traitement est d'autant plus efficace si le site traité possède une couche imperméable en surface pour minimiser les pertes de circulation d'air ou de vapeur. Enfin, l'efficacité des systèmes d'extraction en utilisant de la vapeur d'eau peut être améliorée par une co-injection d'air chaud, mais également par l'utilisation en parallèle de micro-ondes pour chauffer les polluants, augmentant ainsi leur désorption des sols (George *et al.*, 1992).

L'utilisation de liquides (comme de l'eau, parfois combinée avec des solvants) afin de solubiliser les polluants est également une technique de réhabilitation des sols efficace (Ahn



<u>Figure 4 :</u> Principe du lavage *in situ* des polluants du sol.
Une solution de lavage est injectée au niveau de la zone polluée. Cette solution solubilise les contaminants et les transporte jusqu'à la surface où ils seront pompés, traités et éliminés. La solution de lavage peut alors, dans certains cas, être réinjectée pour un nouveau cycle (adapté de Khan *et al.*, 2004).

et al., 2008; Khalladi *et al.*, 2009). Le lavage des sols peut être réalisé *ex situ*, permettant la séparation des solides selon leur taille, leur densité et leur surface chimique. Le lavage des sols *in situ*, quant à lui, se fait uniquement à l'aide de solutions (la plupart du temps, de l'eau additionné de divers surfactants comme le Tween 80, ou bien de certains solvants comme de l'éthanol ou de l'acétone), permettant la solubilisation des polluants par circulation des liquides (Ahn *et al.*, 2008). Les liquides sont ensuite pompés, traités et réutilisés, entraînant ainsi la désorption progressive des polluants des sols (Figure 4). Cependant, l'efficacité de cette technique est inversement proportionnelle à la concentration totale en polluants et ne permet donc pas une dépollution complète. Il est également possible d'installer des structures souterraines (tels que des barrières réactives passives, ou plus simplement des réservoirs souterrains ou des pompes) pour capter et éliminer les polluants contenus dans les liquides (Khan *et al.*, 2004).

Enfin, une autre technique physique de remédiation implique la mobilisation et la migration des polluants dans un champ électrique (par l'usage d'électrophorèse, d'électro-osmose ou d'électrolyse) (Park *et al.*, 2009). Les polluants sont alors collectés et traités. L'électro-remédiation a d'abord été appliquée pour l'extraction de composés ioniques (comme les métaux et les anions inorganiques comme les sulfates), mais peut aussi être efficace sur des huiles ou des graisses (Park *et al.*, 2009). En effet, lors de cette étude récente, l'approche utilisée a permis d'éliminer de 45,1 à 55 % des huiles et graisses présentes, et ce après 17 jours de traitement, au sein d'un sol provenant d'une gare ferroviaire.

4.1.2. Techniques chimiques

Les traitements chimiques sont le plus souvent utilisés pour dépolluer des eaux souterraines. Cependant, les sols (selon leur composition et leur taux d'humidité) peuvent aussi être traitées par cette approche. Les deux techniques les plus communément mises en œuvre sont l'oxydation (ou la réduction), et l'extraction des polluants (regroupant l'hydrolyse, la solubilisation et la manipulation du pH ainsi que l'oxydation et la réduction précédemment citées) (Mulligan *et al.*, 2001; Gan *et al.*, 2009). Ces techniques sont très souvent utilisées en association avec des techniques physiques, comme le lavage des sols ou l'injection de vapeur ou d'eau. L'inconvénient majeur des techniques chimiques est l'introduction de molécules au sein des écosystèmes pouvant engendrer une pollution secondaire. De plus, certains types de polluants ne peuvent être traités de cette façon car ils ont une structure chimique trop proche des composés organiques des sols, comme certains polluants aromatiques, tels que les HAP.

L'oxydation chimique peut permettre une altération des polluants organiques. Dans ce cas, les oxydants chimiques les plus communément utilisés sont le réactif de Fenton (fer ferreux, plus peroxyde d'hydrogène) et l'ozone. Le péroxyde d'azote, le permanganate de potassium, le péroxyde d'hydrogène et le sodium persulfate peuvent également s'avérer efficaces (Gan *et al.*, 2009). Une étude a par exemple démontré que l'ozone généré par un champ électrique permettait de diminuer de plus de 50 % la concentration en phénanthrène dans les sols soumis à 6 heures de traitement (O'Mahony *et al.*, 2006). Néanmoins, ces procédés ne sont pas sans conséquences sur les terres traitées. En effet, les oxydations affectent la mobilité des métaux, la structure de la matière organique et les microorganismes endogènes (Ahn *et al.*, 2005).

L'extraction, quant à elle, peut se faire en utilisant différentes solutions (solvants, surfactants, …) solubilisant les polluants, et permettant leur traitement ultérieur (Viglianti *et al.*, 2006; Ahn *et al.*, 2008). L'ajout de surfactants peut, par exemple, permettre une meilleure élimination des polluants organiques. Une étude récente a ainsi permis de comparer l'efficacité de divers surfactants non ioniques sur un sol pollué par du phénanthrène (Ahn *et al.*, 2008). Cette étude a ainsi démontré que le Brij 30 à 2g/L de concentration permettait d'éliminer jusqu'à 84,1 % de phénanthrène durant le lavage des sols, contrairement au Tween 80, qui, dans les mêmes conditions, ne permet d'éliminer que 72,4 % du phénanthrène. Le défaut principal de ces surfactants est leur adsorption aux particules du sol, pouvant engendrer des pollutions après traitement si tout n'est pas éliminé. Les cyclodextrines, contrairement aux surfactants classiques, sont une alternative non toxique aux solvants et surfactants généralement usités, et sont des molécules de plus en plus utilisées en remédiation (Viglianti *et al.*, 2006). L'utilisation d'huiles végétales est une seconde alternative aux solvants classiques. Elles possèdent ainsi deux avantages principaux : (i) elles permettent une très bonne solubilisation des polluants organiques et (ii), si des traces subsistent au sein de l'écosystème traité, elles seront dégradées par les microorganismes présents (Gan *et al.*, 2009).

4.2. Réhabilitation biologique

Les techniques chimiques et physiques de réhabilitation mises en place afin de décontaminer les sites pollués peuvent être onéreuses et très invasives pour les écosystèmes traités. En effet, réalisées *ex situ*, elles nécessitent l'excavation des écosystèmes à traiter et, réalisées *in situ*, elles requièrent la mise en place de traitement plus ou moins complexes. Cependant, des techniques innovantes et alternatives voient le jour depuis plusieurs années.

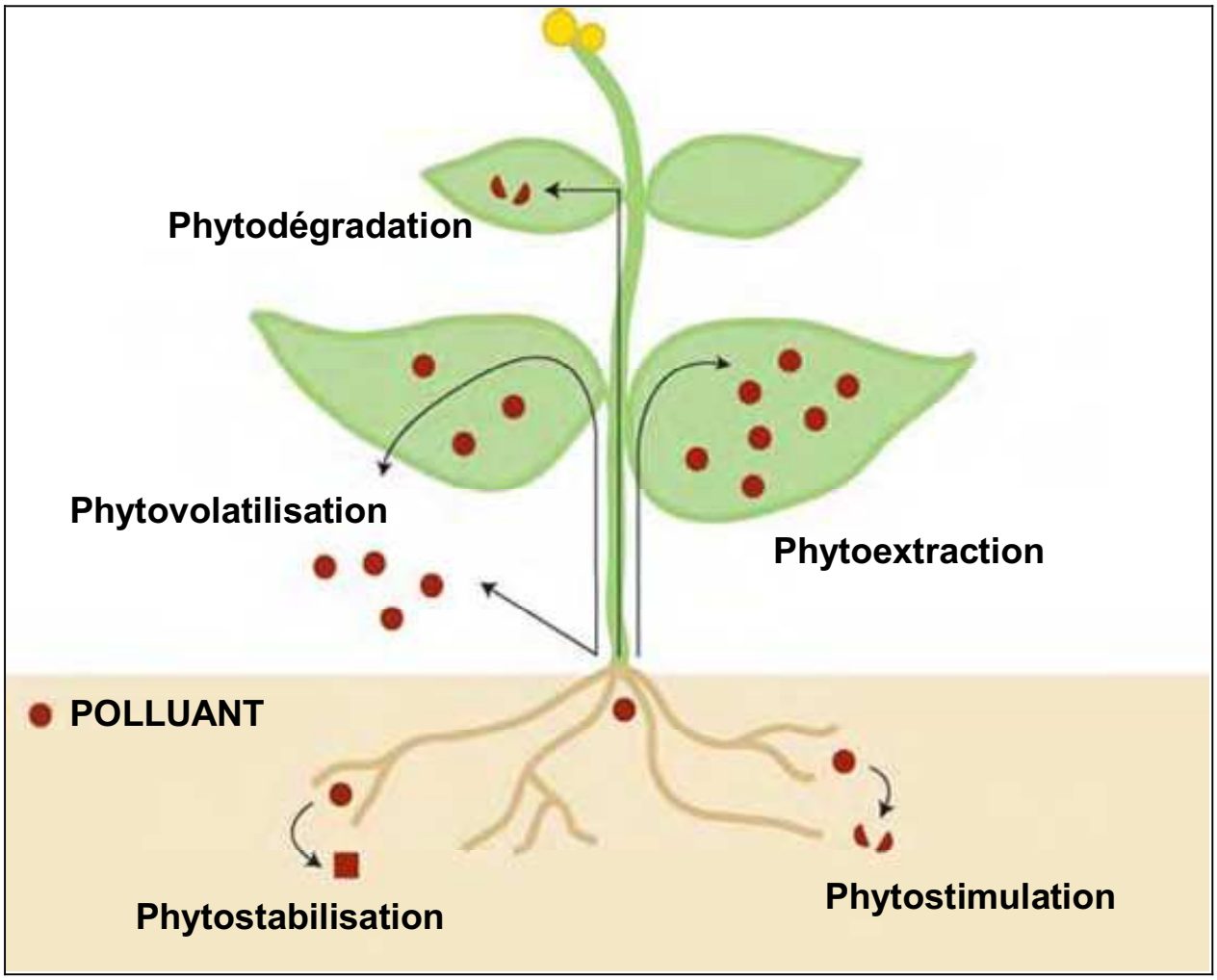


Figure 5 : Principe de la phytoremédiation.
Les polluants (représentés par des cercles rouges) peuvent subir une phytodégradation par la plante, une phytoextraction en étant séquestrés au sein de la partie supérieure de la plante, une phytovolatilisation réalisée par les organes supérieurs de la plante. Au niveau du sol, ils peuvent être dégradés par les microorganismes qui ont été phytostimulés ou être également stabilisés au niveau des racines de la plante, c'est la phytostabilisation (adapté de Pilon-Smits, 2005).

En effet, la biodégradation étant la voie naturelle de recyclage et de réutilisation des déchets en les transformant en nutriments, de nombreuses études visant à appréhender et à stimuler ces voies naturelles de biodégradation ont été entreprises (Gan *et al.*, 2009). La réhabilitation biologique consiste donc à utiliser les capacités naturelles des plantes et des microorganismes à dépolluer les sols (Pilon-Smits, 2005; Andreoni et Gianfreda, 2007).

4.2.1. La phytoremédiation

La phytoremédiation consiste en l'utilisation des plantes et des microorganismes qui leur sont associés pour extraire, réduire et immobiliser les polluants des écosystèmes contaminés, qu'ils soient terrestres, aquatiques ou aériens (Pilon-Smits, 2005). Certaines plantes sont en effet connues pour améliorer la remédiation des sols, via divers processus biochimiques et physiques, comme par exemple la sécrétion de molécules comparables à des « surfactants », permettant d'augmenter la biodisponibilité des polluants (Gan *et al.*, 2009). La phytoremédiation peut donc être utilisée pour traiter de manière efficace des polluants inorganiques (comme les métaux lourds et les radionucléides) ou organiques comme les HAP.

Il existe plusieurs modes d'action de la phytoremédiation, qui dépendent du type de polluant considéré (Figure 5). Ce dernier peut ainsi être séquestré dans les tissus aériens de la plante. Cette dernière est alors éliminée, et l'on nomme ce procédé la phytoextraction. Le polluant peut également être stabilisé au niveau des racines de la plante, qui diminuent les effets de l'érosion, ou en le rendant moins biodisponible par précipitation au niveau de la rhizosphère. Ce procédé est appelé la phytostabilisation. Cependant, bien que la plante puisse accumuler des polluants au sein de ses tissus, il est également possible qu'elle les dégrade en utilisant ses propres capacités métaboliques (on nomme ce procédé la phytodégradation), ou qu'elle les rejette dans l'atmosphère sous forme volatile (ce procédé est appelé la phytovolatilisation) (Pilon-Smits, 2005; Gan *et al.*, 2009). Enfin, les plantes peuvent faciliter la dégradation des polluants par les microorganismes présents au sein de la rhizosphère en les rendant plus accessibles ou en activant les microorganismes responsables de leur dégradation (c'est la phytostimulation).

Plusieurs exemples de l'application de ces méthodes avec plus ou moins de succès sur des polluants organiques ont été décrites, Ainsi, Su et Zhu ont démontré que la majorité des polluants organiques (naphtalène, phénanthrène et pyrène) avait été éliminée par évaporation directe du sol, et non par les germes de riz utilisés (*Oryza sativa*) (Pilon-Smits, 2005; Su et Zhu, 2008). Néanmoins, les derniers essais réalisés avec *Cucurbita pepo* montrent que cette plante est capable d'accumuler certains hydrocarbures (comme le phénanthrène ou le pyrène)

dans les feuilles et les racines (Navarro *et al.*, 2009). Cela est principalement dû à la solubilisation des polluants par ajout de caféine dans les sols testés. De plus, Muratova et ses collaborateurs ont récemment étudié la phytoremédiation de sols contaminés avec du phénanthrène par des graminées (*Sorghum bicolor* (L.) Moench) et ont démontré que les enzymes des racines excrétées par la plante (principalement des peroxydases et des tyrosinases) participaient à la dégradation du phénanthrène (Muratova *et al.*, 2009). Néanmoins, l'attaque initiale des cycles aromatiques avait été réalisée par les microorganismes présents au sein de la rhizosphère, montrant la nécessité de la flore microbienne rhizosphérique.

Bien que la phytoremédiation montre des résultats encourageants, elle présente toutefois des inconvénients non négligeables. Tout d'abord, la plante doit être capable de croître sur le site de la pollution, sachant que ce développement est étroitement lié au niveau de toxicité rencontré mais également à d'autres facteurs externes limitants, comme le climat. La décontamination est également limitée par la taille du système racinaire de l'espèce utilisée (50 cm pour les herbacées et 3 m en moyenne pour les arbres). Un autre inconvénient majeur est le temps nécessaire pour dépolluer les sols *via* la phytoremédiation, ce procédé pouvant prendre des années par rapport à des méthodes comme l'excavation ou l'incinération des sols. Enfin, une contamination potentielle de la chaîne alimentaire *via* une consommation accidentelle des plantes épuratrices au niveau des sites traités est également un risque important à prendre en considération (Pilon-Smits, 2005; Gan *et al.*, 2009).

4.2.2. La bioremédiation microbienne

Le principe de bioremédiation microbienne repose sur l'utilisation des capacités intrinsèques des microorganismes à métaboliser des polluants, et est appliqué de façon effective depuis maintenant plusieurs années. La bioremédiation est généralement réalisée par des groupes de souches appelés *consortia* et contrairement aux traitements classiques de remédiation, c'est un procédé principalement *in situ*, donc moins invasif, qui représente par conséquent une nouvelle alternative de traitement des environnements pollués (Thouand *et al.*, 1999; Gan *et al.*, 2009). Il existe aujourd'hui plusieurs approches de bioremédiation : celles qui utilisent directement les capacités métaboliques des microorganismes présents dans les environnements pollués, et qui correspondent à l'atténuation naturelle et à la biostimulation, et celles qui reposent sur l'introduction dans ces sols de microorganismes présentant les capacités métaboliques nécessaires à la dégradation des polluants présents, que l'on nomme la bioaugmentation (Andreoni et Gianfreda, 2007).

L'atténuation naturelle est l'une des premières techniques à avoir été exploitée. Elle consiste à laisser se développer les bactéries endogènes présentes dans les sols pollués, afin qu'elles utilisent leurs capacités métaboliques naturelles pour assurer la biodégradation des composés xénobiotiques sans intervention humaine. Cette stratégie de décontamination est peu onéreuse, mais elle a le principal désavantage d'être relativement longue (Mulligan et Yong, 2004) par rapport à d'autres techniques physiques, chimiques ou biologiques. De plus, des procédés abiotiques, tels que l'évaporation, la dissolution, la dispersion, l'émulsification, l'adsorption ou la photooxydation, peuvent aussi entrer en jeu dans l'élimination des polluants considérés (Mulligan et Yong, 2004). Une étude réalisée en 2001 par Margesin et Schinner a montré une dégradation de 50 % des hydrocarbures après 3 ans sur un site pollué par du diesel (2 600 mg de diesel/kg sol), à une altitude de 2 875 m (Margesin et Schinner, 2001).

L'atténuation naturelle cède de plus en plus sa place à une autre stratégie : la biostimulation. Ce procédé consiste à modifier les conditions environnementales par adjonction de nutriments, d'eau, de donneurs ou d'accepteurs d'électrons, afin d'améliorer les métabolismes épurateurs, et donc d'accélérer les processus naturels de biodégradation (Andreoni et Gianfreda, 2007). Ces ajouts permettent d'améliorer la biodisponibilité du polluant et/ou de modifier l'environnement oxydo-réducteur de l'écosystème (la disponibilité en accepteurs ou donneurs d'électrons peut influencer les activités microbiennes) et/ou la biodisponibilité de divers nutriments (par exemple, pour aider au développement des communautés présentes). Une étude récente réalisée *in situ* sur un suivi de 90 jours démontre l'efficacité de l'apport d'amidon, de glucose et de succinate de sodium dans l'accélération de la dégradation du phénanthrène et du benzo(*a*)pyrène (Teng *et al.*, 2009). Cependant, ces techniques dépendent de nombreux paramètres difficilement contrôlables comme l'hétérogénéité des sols, la présence de microorganismes compétiteurs, ou encore la présence de nutriments plus facilement utilisables (Gan *et al.*, 2009).

Parfois, les populations microbiennes endogènes d'un écosystème pollué ne possèdent pas les capacités de dégradation des polluants ou bien alors ces capacités sont peu efficaces, voire partielles. C'est pourquoi, pour pallier cette déficience métabolique, des procédés de bioaugmentation voient le jour (Thouand *et al.*, 1999; Andreoni et Gianfreda, 2007). Ils consistent à ensemencer le sol avec des microorganismes adaptés à la dégradation du (ou des) xénobiotique(s) présent(s), afin d'assurer ou d'accélérer le processus de dépollution. Cet ajout peut se faire sous la forme d'une souche unique, ou d'un groupe de souches appelé *consortia* (Samanta *et al.*, 2002; Silva *et al.*, 2009; Zanaroli *et al.*, 2010). Une étude intéressante de bioremédiation (associant l'apport de nutriments et de microorganismes), effectuée *in situ* sur

une plage d'Israël contaminée par des hydrocarbures a été réalisée, après optimisation de conditions par des études préliminaires (Rosenberg *et al.*, 1992). Cette étude a démontré la diminution en 28 jours de la concentration en polluants, qui est passée de 5,1mg/g de sable à 0,6mg/g de sable (entraînant une baisse de 88% de la contamination, le sol non traité ne subissant qu'une baisse effective de 15%).

Ce type d'approche présente plusieurs problèmes majeurs. Ainsi, les souches introduites peuvent rapidement décliner suite aux stress biotiques et/ou abiotiques générés par le nouvel environnement. Les souches exogènes ont donc des difficultés à s'implanter efficacement au sein de l'écosystème à dépolluer. Par exemple, il peut arriver que les souches microbiennes introduites limitent le développement des populations endogènes. Ainsi, une étude réalisée sur des microcosmes de sols pollués par du phénanthrène démontre que l'introduction de la souche *Sphingomonas paucimobilis* 20006FA (isolée d'une environnement pollué pour ses capacités d'utilisation du phénanthrène et de l'anthracène comme seule source de carbone et d'énergie) n'améliore pas significativement la dégradation de ce dernier (Coppotelli *et al.*, 2008). Cela semble lié à l'accumulation de certains métabolites toxiques produits durant la dégradation des HAP par cette souche 20006FA, qui entraîne le déclin des communautés dépolluantes endogènes, diminuant ainsi l'efficacité de dégradation des HAP présents. Les résultats de cette étude démontrent un impact à long terme de la souche introduite sur la diversité de la communauté bactérienne initialement présente. En effet, une forte réduction de la diversité est visible, et est la conséquence de l'introduction de la souche 20006FA. Cet impact serait donc attribué à la pression de sélection engendrée par la présence et l'accumulation de métabolites de dégradation du phénanthrène.

A plus long terme, la bioaugmentation pourrait aussi impliquer des organismes génétiquement modifiés (OGM) particulièrement bien adaptés pour une pollution donnée. L'utilisation de tels microorganismes *in situ* soulève cependant le problème de leur dissémination dans l'environnement, et du transfert éventuel de matériel génétique vers d'autres microorganismes. Par exemple, la sélection des organismes génétiquement modifiés doit utiliser un gène qui ne peut se disséminer (comme un gène de résistance à un antibiotique), mais plutôt de gènes de résistance à des métaux ou à des molécules comme certains xénobiotiques. Pour tenter de s'affranchir de ces inconvénients majeurs, des systèmes de suivis et de contrôles des microorganismes dans le sol sont actuellement à l'étude. Il s'agit le plus souvent d'un système suicide éliminant spontanément les microorganismes du site une fois le polluant épuisé (Davison, 2002).

5. Conclusion

Les procédés de bioremédiation sont donc des alternatives prometteuses par rapport aux techniques classiques de dépollution. En effet, ils offrent la possibilité de réaliser une décontamination d'environnements pollués en diminuant les impacts sur les écosystèmes traités et à des coûts très réduits. Afin d'améliorer l'efficacité de ces procédés de dépollution, il est néanmoins nécessaire d'identifier les microorganismes épurateurs et de caractériser leurs capacités de dégradation. La diversité microbienne des écosystèmes étant considérable, cette tâche reste difficile. Lier la structure à la fonction demeure un des enjeux majeurs de l'écologie microbienne. Concernant les HAP, comme nous le verrons dans le prochain chapitre, les voies de dégradation sont complexes et peuvent être portées par un nombre important de microorganismes.

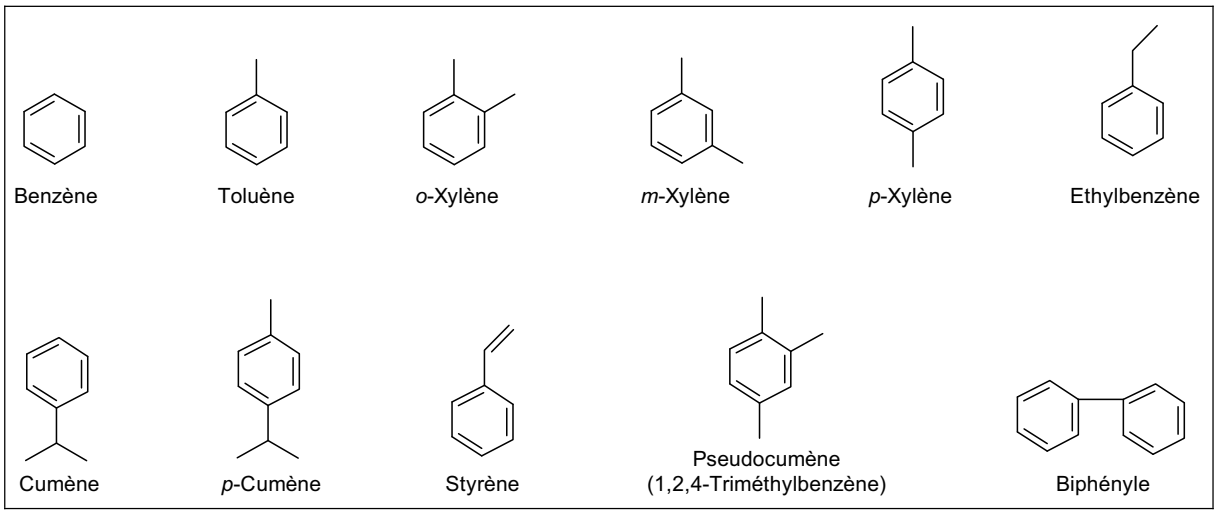
Chapitre II : Biodégradation des HAP par les microorganismes

1. Introduction

Les hydrocarbures aromatiques sont des composés que l'on retrouve de manière naturelle au sein des environnements (produits pétrolifères, dégradation de la matière organique comme la lignine, feux de forêts, activités volcaniques,…) (Vandecasteele, 2005). Cependant, de nombreuses activités anthropiques (comme le raffinage du pétrole brut ou l'expansion des industries chimiques) entraînent une forte production de ces molécules, qui malheureusement dans certains cas, vont s'accumuler au sein des sols et donc générer des pollutions. La prise de conscience des problèmes posés par les sols pollués, notamment au regard de leur impact sur la santé, les aquifères et la ressource en eau, a contribué à promouvoir l'étude des processus de biodégradation dans les différentes situations aérobies et anaérobies envisageables.

2. Les hydrocarbures aromatiques

Les hydrocarbures aromatiques sont des composés toxiques, de par leur caractère mutagène, carcinogène et récalcitrant (Phale *et al.*, 2007). Ils constituent une classe de composés hydrophobes qui obéissent à la règle de Hückel. Cette règle peut s'exprimer ainsi : « un hydrocarbure est aromatique s'il est plan, et s'il possède 4n + 2 électrons délocalisables dans un système cyclique (où n est un entier naturel qui correspond au nombre de cycles) ». A titre d'exemple, le benzène, qui est le membre le plus simple des hydrocarbures aromatiques, a une composition élémentaire de C_6H_6 et est plan. Il est composé d'un cycle à 6 côtés et possède trois doubles liaisons. Les électrons associés à ces doubles liaisons sont délocalisables, lui donnant une stabilité thermodynamique non négligeable. Leur stabilité (fortement dépendante de l'agencement des cycles), leur faible solubilité et volatilité font que ces hydrocarbures aromatiques sont persistants dans l'environnement et entraînent des pollutions des écosystèmes (Haritash et Kaushik, 2009). Les hydrocarbures aromatiques sont séparés en deux groupes distincts : les hydrocarbures monoaromatiques et les hydrocarbures aromatiques polycycliques.



<u>Figure 6 :</u> Structure et nomenclature d'hydrocarbures monoaromatiques.
Seul ceux les plus fréquemment rencontrés, du fait de leur abondance dans les essences et de leurs utilisations en pétrochimie, sont représentés (tirée de Vandecasteele, 2005).

2.1. Les hydrocarbures monoaromatiques

Les hydrocarbures monoaromatiques forment une des grandes classes de composés aromatiques présents dans l'environnement avec les hydrocarbures chloroaromatiques. La production naturelle principale de composés monoaromatiques provient de la dégradation de la lignine (Vandecasteele, 2005). Ces hydrocarbures monoaromatiques peuvent aussi avoir une origine biosynthétique. C'est le cas par exemple, du *p*-cymène (ou *p*-isopropyltoluène), un composant des huiles essentielles. C'est aussi celui du styrène, présent dans divers produits naturels comme certains fruits, ou encore certaines plantes. Cependant, ce dernier est aussi un produit pétrochimique de gros tonnage et représente un polluant important (Phale *et al.*, 2007). En fait, les quantités considérables d'hydrocarbures monoaromatiques rencontrées dans l'environnement sont en grande majorité d'origine pétrolière, ou pétrochimique, et sont la conséquence des pollutions très nombreuses engendrées par l'activité humaine (Andreoni et Gianfreda, 2007). L'existence d'hydrocarbures pétroliers dans l'environnement n'est cependant pas uniquement liée aux activités humaines. En effet, cette présence peut être très ancienne et résulter des suintements naturels des sources naturelles pétrolières.

Les structures de certains hydrocarbures monoaromatiques sont présentées dans la Figure 6. Parmi ceux-ci, le benzène, le toluène, l'éthylbenzène et les xylènes, désignés sous le sigle de BTEX, sont particulièrement importants du fait de leur abondance dans les essences, et de leurs utilisations en pétrochimie (Phale *et al.*, 2007). Ces composés ont une solubilité dans l'eau importante, et une volatilité notable qui décroissent respectivement avec le nombre d'atomes de carbone (Andreoni et Gianfreda, 2007). Ils présentent par ailleurs des valeurs d'hydrophobicité (log $P_{o/w}$) entre 2 et 3. Ces propriétés ont une incidence directe, et de première importance sur leurs caractéristiques biologiques. Leur solubilité élevée leur confère ainsi une mobilité importante, dans les eaux et les sols (transfert et migration dans les aquifères) et une accessibilité importante pour les microorganismes. Cependant, leur hydrophobicité correspond à des valeurs de log $P_{o/w}$ de forte toxicité pour la membrane cytoplasmique (ils vont ainsi modifier la structure de la membrane et diminuer son efficacité). Les hydrocarbures monoaromatiques sont ainsi toxiques pour les cellules eucaryotes.

2.2. Les hydrocarbures aromatiques polycycliques (HAP)

2.2.1. Structures, nomenclature et propriétés physico-chimiques

Les HAP sont constitués d'au moins deux cycles aromatiques benzéniques fusionnés. Au sens strict, ils ne contiennent donc que des atomes de carbone et d'hydrogène (Vandecasteele, 2005; Seo *et al.*, 2009). Les structures des 16 HAP non substitués retenus

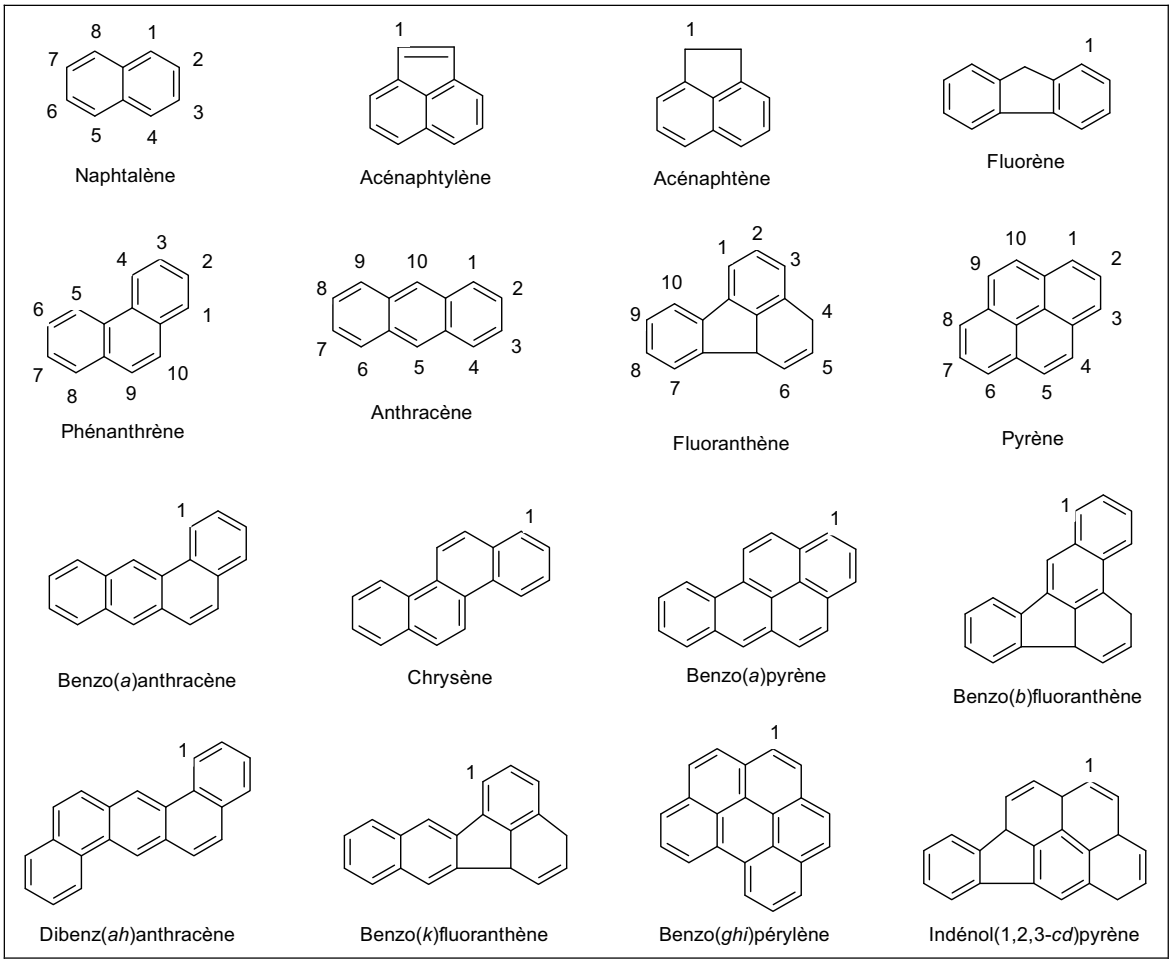


Figure 7 : Structure et nomenclature des seize HAP prioritaires de la liste EPA.
Les atomes de carbone extérieurs sont numérotés dans le sens anti-trigonométrique avec l'origine sur le sommet libre le plus à gauche du cycle supérieur le plus à droite (exemples pour le naphtalène, le fluoranthène et le pyrène). L'origine est différente selon que le rang supérieur comporte un cycle unique (cas du fluoranthène) ou plusieurs cycles (cas du pyrène). Le phénanthrène et l'anthracène sont des exceptions à ces règles, la notation indiquée étant celle utilisée dans le domaine de la biodégradation (tirée de Vandecasteele, 2005).

Tableau 1 : Propriétés physico-chimiques de plusieurs HAP.
La solubilité est déterminée dans l'eau à 25°C. Le Log $P_{o/w}$ est défini comme le logarithme décimal du coefficient de partition Kow d'un composé donné dans un système standard diphasique octanol/eau (tiré de Vandecasteele, 2005).

HAP	Masse molaire (g/mol)	Point de fusion (°C)	Point d'ébullition (°C)	Solubilité (mg/L)	Log $P_{o/w}$
Naphtalène	128,2	80	218	32	3,35
Acénaphtylène	152,2	82	270	3,93	
Acénaphtène	154,2	93	279	3,42	3,92
Fluorène	166,2	116	294	1,9	4,18
Phénanthrène	178,2	100	338	1,0	4,52
Anthracène	178,2	216	340	0,07	5,54
Fluoranthène	202	107	383	0,27	5,22
Pyrène	202	150	393	0,16	5,18
Chrysène	228,2	254	441	0,006	5,79
Benzo(*a*)pyrène	252	179	496	0,0038	5,98

comme polluants prioritaires par l'*American Environmental Protection Agency* (EPA) sont présentés dans la Figure 7. Depuis quelques années, certains de ces 16 HAP font également partie des listes de l'OMS (Organisation Mondiale de la Santé) et de la communauté européenne. Ces HAP sont constitués de cycles aromatiques accolés en nombre croissant, allant de deux cycles (comme le naphtalène) et n'ayant pas de limite supérieure encore démontrée. L'agencement des cycles peut être linéaire (anthracène), angulaire (fluoranthène) ou groupé (pyrène). Parmi ces HAP, on fait parfois une distinction entre ceux à bas poids moléculaire (deux et trois cycles), et ceux à haut poids moléculaire (quatre cycles et plus). Cependant, certains composés aromatiques contenant du soufre (comme par exemple le phénanthrothiophène : $C_{14}H_8S$), de l'azote (comme la quinoline : C_9H_7N) ou de l'oxygène (comme le dibenzofurane : $C_{12}H_8O$) sont parfois associés, et ces composés peuvent alors être aussi considérés comme des HAP.

Les HAP non substitués sont très majoritairement retrouvées chez les HAP d'origine pyrolytiques. Au niveau des produits pétroliers par contre, ce sont les HAP présentant des substituants alkyles (notamment méthyles), en différentes positions qui sont majoritaires. Ces structures vont définir un certain nombre de propriétés physico-chimiques, présentées dans le Tableau 1. La stabilité des HAP est grandement fonction de l'arrangement des cycles, les HAP angulaires étant les plus stables, et les linéaires les moins stables(Vandecasteele, 2005; Seo *et al.*, 2009). Leur solubilité en milieu aqueux est notable pour le naphtalène (32 mg/L avec deux cycles aromatiques), mais décroît rapidement avec le nombre de cycles aromatiques (4 µg/L pour le benzo(*a*)pyrène qui se compose de cinq cycles). Il en va de même pour leur volatilité. De plus, comme détaillé dans le Chapitre I, les HAP sont adsorbables sur divers supports solides, ce qui leur confère une propriété importante pour leur devenir dans l'environnement.

2.2.2. Origines pyrolytiques et pétrogéniques des HAP

Les HAP retrouvés dans l'environnement sont essentiellement d'origine pyrolytiques et pétrogéniques. On estime par exemple que 2,3 Mt/an de pétrole ont été rejetées dans l'ensemble des océans du globe, regroupant les contaminations liées aux activités anthropiques, mais également les suintements naturels de gisements (Vandecasteele, 2005). L'émission atmosphérique de HAP (gaz automobiles, émissions domestiques et industrielles par exemple) est également une source, et un facteur de dissémination dans l'environnement (Tableau 2 pages suivante) (Seo *et al.*, 2009).

Tableau 2 : HAP prédominants dans les émissions atmosphériques de diverses sources.
Un + indique la présence du HAP au sein des émissions atmosphériques analysées.
Source : http://www.ineris.fr/

	Chauffage domestique	**Véhicule à essence**	**Véhicule diesel**	**Raffinerie de pétrole**	**Centrale électrique à charbon**
Fluorène				+	+
Phénanthrène			+	+	+
Anthracène				+	
Fluoranthène	+		+		+
Pyrène	+		+	+	
Benzo(*a*)anthracène	+				
Chrysène	+				
Benzo(*a*)pyrène		+			
Indéno(1,2,3-*cd*)pyrène		+			
Coronène		+			

Actuellement, c'est l'origine pyrolytique anthropique qui est considérée comme la source majeure de HAP dans l'environnement, notamment à cause des émissions domestiques et industrielles (Haritash et Kaushik, 2009). Ces dernières induisent la formation de produits liquides riches en divers HAP, comme le créosote (pouvant contenir plus de 30 HAP différents et dont la concentration totale peut atteindre 85% du produit) et l'huile d'anthracène (contenant de l'anthracène à 95%, mélangé à d'autres HAP comme le phénanthrène), mais aussi l'émission de gaz contenant des hydrocarbures et leurs dérivés oxygénés : du monoxyde de carbone et des oxydes d'azote. Les HAP d'origine pyrolytique sont caractérisés par la prédominance des HAP non substitués (sans groupement alkyles), sur leurs homologues alkylés beaucoup moins stables (avec un ou plusieurs groupements alkyles comme le méthyle par exemple). Cette différence est liée à la température de combustion à laquelle se forment les composés. En effet, les HAP pyrolytiques sont générés par des processus de combustion incomplète de la matière organique à haute température, avec une déficience en oxygène durant la combustion.

Les produits pétroliers contenant des HAP, principalement alkylés, correspondent au pétrole brut, mais également aux produits de raffinage (hormis les fractions légères) comme le kérosène, le gazole, le fioul domestique, les huiles, les fiouls lourds et les bitumes (Haritash et Kaushik, 2009). Ils peuvent contenir des HAP de très haute masse moléculaire (C_{70} et plus), non encore analysés et identifiés individuellement, où la contribution des substituants alkyles peut être très importante. Les HAP pétrogéniques résultent principalement de la formation du pétrole par catagenèse, un processus thermique qui se produit à des températures relativement basses (50-150°C), et qui permet la conservation des chaînes alkylées (Haritash et Kaushik, 2009). Il faut également noter que le rejet d'huiles dans l'environnement donne lieu à une pollution diffuse (échappements, combustion, fuites d'huile), à laquelle s'ajoutent les rejets illégaux d'huiles de vidange. De plus, ces polluants étant sous forme liquides, ils se retrouvent souvent dans le réseau fluvial. Enfin, il faut aussi mentionner les bitumes, épandus en quantités énormes dans l'environnement, non comme rejets, mais comme revêtements de sols, et qui ne sont donc pas considérés comme des polluants (Peng *et al.*, 2008).

3. Microorganismes dégradant les HAP

Les HAP sont des composés très riches en carbone et sont une source d'énergie non négligeable au sein des sols, ce qui a conduit au développement d'organismes dits hydrocarbonoclastes obligatoires (principalement des bactéries et des champignons)

(Fernández-Luqueño *et al.*, 2010). Ces microorganismes sont dépendants de ces sources de carbone et d'énergie pour se développer, et jouent un rôle très important dans la dépollution des environnements contaminés. Ces microorganismes agissent rarement seuls, et sont plus généralement regroupés au sein de *consortia*. Il a ainsi été décrit que certaines exoenzymes de champignons attaquaient les HAP de haut poids moléculaire, les métabolites obtenus étant par la suite dégradés par les communautés microbiennes (Scullion, 2006).

Certains de ces microorganismes épurateurs ont de plus la capacité d'améliorer la solubilisation des HAP, molécules fortement hydrophobes, par la production de biosurfactants (formation de micelles), ou par la production de biofilms (matrices extracellulaires adhésives et protectrices) (Johnsen et Karlson, 2004; Leglize *et al.*, 2008). C'est par exemple le cas de la souche bactérienne *Pseudomonas aeruginosa* P-CG3, capable de produire un biosurfactant améliorant fortement la solubilisation du phénanthrène et du pyrène au sein de sols contaminés (Cheng *et al.*, 2004). La formation de biofilms est également une des stratégies mises en place dans l'amélioration de la dégradation des HAP (Andreoni *et al.*, 2004; Johnsen et Karlson, 2004). Cette formation de biofilms permet à la fois aux microorganismes de se protéger des effets toxiques inhérents aux HAP, mais aussi d'améliorer leur biodisponibilité.

Ces microorganismes ont donc développé diverses capacités leur permettant de métaboliser les HAP. Certaines biodégradations (en conditions aérobies ou anaérobies) peuvent être complètes ou non (Vandecasteele, 2005; Haritash et Kaushik, 2009). Les microorganismes impliqués dans cette biodégradation peuvent être : des bactéries, des archées, des algues ou encore des champignons. On dénombre après un siècle d'études 200 genres de ces microorganismes, représentant plus de 500 espèces et souches décrites (Yakimov *et al.*, 2007).

3.1. Les microorganismes anaérobies

Le catabolisme anaérobie est un processus très important de la biodégradation de nombreux composés dans l'environnement. Les premières études ont été réalisées par Mihelcic et Luthy en 1988, qui ont pu mettre en évidence la minéralisation du naphtalène et de l'acénaphtène en conditions anaérobies (Mihelcic et Luthy, 1988a, b). La majorité des études ont été réalisées sur des microcosmes inoculés avec des sédiments marins. Cependant, des résultats intéressants ont également été obtenus avec des *inocula* d'eau douce (Meckenstock *et al.*, 2004). Ces dernières années plusieurs études ont permis de caractériser, et dans certains cas d'isoler de nouvelles souches de l'environnement (bactéries, archées, champignons) capables de dégrader divers HAP en conditions anaérobies. Par exemple, en

<u>Tableau 3 :</u> Souches bactériennes isolées de l'environnement capables de dégrader divers HAP (liste incomplète).

NAP : naphtalène ; FLE : fluorène ; PHE : phénanthrène ; FLA : fluoranthène ; ANT, anthracène ; PYR : pyrène ; MNAP : méthyl-naphtalène ; BaP : Benzo(*a*)pyrène ; BaA : benzo(*a*)anthracène ; dMBaA : diméthylbenzo(*a*)anthracène ; DBA : dibenzo(*a,h*)anthracène ; COR : coronène; CHR : chrysène (adapté de Seo et al., 2009).

Nom de l'espèce	Nom de la souche	HAP dégradé(s)
Alcaligenes denitrificans		FLA
Arthrobacter sp.	F101	FLE
Arthrobacter sp.	P1-1	PHE
Arthrobacter sulphureus	RKJ4	PHE
Acidovorax delafieldii	P4-1	PHE
Bacillus cereus	P21	PYR
Brevibacterium sp.	HL4	PHE
Burkholderia sp.	S3702, RP007, 2A-12TNFYE-5, BS3770, C3	PHE
Burkholderia cepacia	BU-3	NAP, PHE, PYR
Burkholderia cocovenenans	LB400	PHE
Cycloclasticus sp.	P1	PYR
Janibacter sp.	YY-1	FLE, PHE, ANT
Marinobacter	NCE312	NAP
Mycobacterium sp.		PYR, BaP
Mycobacterium sp.	JS14	FLA
Mycobacterium sp.	6PY1, KR2, AP1, KMS	PYR
Mycobacterium sp.	RJGII-135	PYR,BaA, BaP
Mycobacterium sp.	PYR-1, LB501T	FLA, PYR, PHE, ANT
Mycobacterium sp.	CH1, BG1, BB1, KR20	PHE, FLE, FLA, PYR
Mycobacterium flavescens		PYR, FLA
Mycobacterium vanbaalenii	PYR-1	PHE, PYR, dMBaA
Nocardioides sp.	KP7	NAP, PHE
Pasteurella sp.	IFA	FLA
Polaromonas naphthalenivorans	CJ2	NAP
Pseudomonas sp.	C18, PP2, DLC-P11	NAP, PHE
Pseudomonas sp.	NCIB 9816-4, F274	FLE
Pseudomonas paucimobilis		PHE
Pseudomonas vesicularis	OUS82	FLE
Pseudomonas putida	P16, BS3701, BS3750, BS590-P, BS202-P1	NAP, PHE
Pseudomonas putida	CSV86	MNAP
Pseudomonas fluorescens	BS3760	PHE, CHR, BaA
Pseudomonas stutzeri	P15	PYR
Pseudomonas saccharophilia		PYR
Pseudomonas aeruginosa		PHE
Ralstonia sp.	U2	NAP
Rhodanobacter sp.	BPC-1	BaP
Rhodococcus sp.		PYR, FLA
Staphylococcus sp.	PN/Y	PHE
Stenotrophomonas maltophilia	VUN 10,010	PYR, FLA, BaP
Stenotrophomonas maltophilia	VUN 10,003	PYR, FLA, BaA, BaP, DBA, COR
Sphingomonas yanoikuyae	R1	PYR
Sphingomonas yanoikuyae	JAR02	BaP
Sphingomonas sp.	P2, LB126	FLE, PHE, FLA, ANT
Sphingomonas paucimobilis	EPA505	FLA, NAP, ANT, PHE
Terrabacter sp.	DBF63	FLE
Xanthamonas sp.		PYR

2008, Chang et ses collaborateurs ont étudié la dégradation anaérobie du phénanthrène et du pyrène au sein de sédiments de mangrove (Chang *et al.*, 2008). La souche ayant la plus grande capacité de biodégradation est la souche MSA3, qui semble proche de l'espèce *Clostridium pascui*, connue comme dégradant divers HAP en conditions anaérobies (Yuan et Chang, 2007). Au cours d'une autre étude, toujours au sein de sédiments de mangrove contaminés par des HAP, plusieurs souches se développant en conditions pauvres en oxygène (0,2 % d'oxygène) et/ou sans oxygène ont été isolées (Li *et al.*, 2009a). Deux souches étaient affiliées au genre *Sphingomonas,* une à *Microbacterium* et une à *Rhodococcus*. L'analyse de ces souches a mis en évidence leur capacité de biodégrader divers HAP comme le fluorène, le phénanthrène, le fluoranthène ou encore le pyrène. D'autres études ont aussi porté sur des souches déjà connues, mais où les capacités de biodégradation anaérobie n'avaient pas encore été démontrées. C'est notamment le cas de l'étude menée sur *Bacillus cereus,* qui, en conditions anaérobies, et avec ajout de surfactants anioniques, est capable de métaboliser le fluoranthène (Fuchedzhieva *et al.*, 2008).

Les archées sont également impliquées dans la biodégradation anaérobie des HAP (Chang *et al.*, 2006; Kim *et al.*, 2008). Les espèces dominantes retrouvées dans des sédiments de la baie de Baltimore, après addition de phénanthrène sont *Methanosarcina semesiae* et *Methanosarcina lacustris* et semblent impliquées dans la biodégradation anaérobie des polluants (Chang *et al.*, 2006). En effet, lors de l'introduction d'un inhibiteur de la méthanogénèse (l'acide bromoéthanosulfonique), plus aucune dégradation des HAP n'est observée, montrant l'implication indirecte des communautés d'archées méthanogènes dans la biodégradation de ces polluants, car la méthanogénèse semble indispensable. Une autre étude sur des sédiments marins de la baie de Gwangyang, en Corée, semble également montrer une implication des archées dans la dégradation anaérobie des HAP (Kim *et al.*, 2008). Plusieurs séquences semblent s'affilier à l'ordre des *Methanosarcinales* alors que d'autres sont apparentées au *phyla* des *Crenarchaeota*.

3.2. Les microorganismes aérobies

Les bactéries constituent les microorganismes les plus étudiés et de nombreuses souches possédant la capacité d'utiliser les HAP comme seule source de carbone et d'énergie ont été isolées et étudiées (Seo *et al.*, 2009). Ces souches peuvent appartenir à plusieurs phyla, classes ou genres bactériens (Tableau 3). Les plus souvent rencontrés sont les genres *Mycobacterium*, *Rhodococcus*, *Nocardioides*, *Pseudomonas*, *Alcaligenes*, ou *Sphingomonas.* Ces souches ont été généralement isolées de sols, ou de sédiments d'eau douce. Cependant, de

plus en plus d'études s'intéressent également aux environnements marins (Haritash et Kaushik, 2009). Les espèces du genre *Pseudomonas* ont été les premières décrites comme possédant des capacités métaboliques pour assurer la dégradation des HAP en condition aérobie (Fernley et Evans, 1958). Les bactéries utilisant les HAP à quatre cycles (ainsi que le fluorène) comme seule source de carbone ont été isolées à partir de la fin des années 1980 (Vandecasteele, 2005). La dégradation des HAP à faible nombre de cycle comme le phénanthrène ou le naphtalène font l'objet de nombreuses études, et les voies métaboliques impliquées ont été largement décrites et seront détaillées dans le prochain paragraphe (Peng *et al.*, 2008). Les études récentes se focalisent à présent sur la caractérisation des souches bactériennes capables de dégrader des HAP à grand nombre de cycles, comme le chrysène (Willison, 2004), et le benzo(*a*)anthracène (Schuler *et al.*, 2009), chacun composé de quatre cycles et qui s'avèrent être les plus récalcitrants à la dégradation.

Les bactéries ne sont pas les seules capables d'assurer la biodégradation des HAP. En effet, certains champignons parmi lesquels il est important de distinguer les champignons filamenteux lignolytiques (comme *Phanerochaete chrysosporium*), des champignons non lignolytiques (*Aspergillus, Cunninghamella, Penicillium,…*) ou des levures (*Candida, Saccharomyces*...) peuvent permettre cette dégradation (en grande majorité par des réactions de cométabolisme) (Cerniglia, 1992). Parmi des centaines d'espèces de champignons ayant des capacités lignolytiques, et donc potentiellement capable de dégrader des HAP, *Phanerochaete chrysosporium*, *Bjerkandera adusta* et *Pleurotus ostreatus* ont été les plus étudiées (Cerniglia, 1997). Il semble ainsi que les capacités de biodégradation soient induites par la présence des HAP. La biodégradation des HAP réalisée par les champignons filamenteux non lignolytiques a été fortement étudiée durant ces dernières années. Par exemple, une étude a pu mettre en évidence treize souches de deuteromycètes isolées du sol capables de dégrader le naphtalène, le phénanthrène et/ou l'anthracène (Clemente *et al.*, 2001). Ces capacités de biodégradation sont principalement dues à la faible spécificité de substrat des enzymes extracellulaires produites par ces microorganismes, comme les peroxydases, les rendant aptes à dégrader divers substrats par co-métabolisme comme les HAP à grand nombre de cycles, pourtant récalcitrants (Haritash et Kaushik, 2009). Finalement, le dernier type de champignon également retrouvé au sein des environnements pollués correspond aux levures. Une étude publiée en 2002 par Gaspar et collaborateurs, porte sur le développement racinaire du champignon *Glomus geosporum* sur les racines de maïs (Gaspar *et al.*, 2002). En présence de phénanthrène, ce développement est limité. Cependant, en présence de la levure *Rhodotorula glutinis*, il y a baisse de l'accumulation de phénanthrène

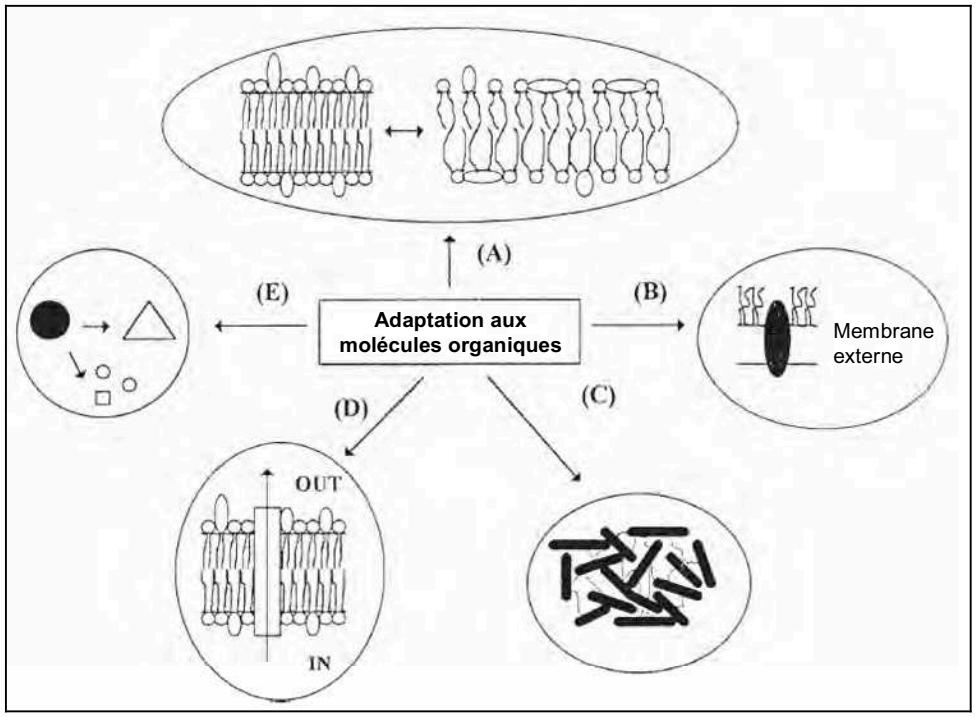
et augmentation de la colonisation par *Glomus geosporum*, traduisant une dégradation potentielle du phénanthrène par la levure *Rhodotorula glutinis*.

Enfin, depuis le début des années 1980, les cyanobactéries et les algues sont également connues pour leur capacité à dégrader partiellement les HAP. Ainsi, de nombreuses études font état de la dégradation complète ou partielle de divers HAP en présence d'algues ou de cyanobactéries (Haritash et Kaushik, 2009). Un exemple récent montre que quatre espèces de microalgues (*Chlorella vulgaris, Scenedesmus platydiscus, Scenedesmus quadricauda,* et *Selenastrum capricornutum*) ont la capacité de dégrader le fluoranthène et/ou le pyrène, avec des efficacités dépendantes de l'espèce et de la toxicité du composé (Lei *et al.*, 2007). De nombreuses connaissances ont été acquises sur les voies de dégradation en aérobiose que nous décrivons dans le paragraphe suivant. Il faut noter que l'efficacité de dégradation des HAP en aérobiose est plus importante que dans des conditions dépourvues d'oxygène.

4. Transport passif et actif des HAP

Le problème de la dégradation des HAP se pose d'abord en termes d'accessibilité pour les microorganismes. En effet, cette accessibilité est fortement diminuée en raison de la très faible solubilité de ces substrats. Pour limiter cette contrainte, différents paramètres sont à prendre en compte, comme la solubilisation dans la phase aqueuse, l'accession interfaciale, l'adsorption des HAP à des surfaces solides, mais aussi le rôle des surfactants dans la solubilisation micellaire des HAP, et dans la diminution des tensions de surface (Isken et de Bont, 1998; Kallimanis *et al.*, 2007).

Dans les conditions optimales, les HAP, rendus accessibles aux microorganismes, doivent généralement traverser la paroi et la membrane cellulaire pour être dégradés. Chez les bactéries Gram-positive, la paroi est formée d'une couche épaisse de peptydoglycanes fortement pontées, dont l'épaisseur peut atteindre jusqu'à 10 fois celle d'une bactérie Gram-négative, et peut donc constituer une barrière de perméabilité. Il n'y a cependant pas d'information disponible sur le transport des HAP au sein des parois bactériennes. Les bactéries Gram-négative possèdent également une barrière de perméabilité correspondant à la membrane externe, qui s'ajoute à la membrane cytoplasmique. Ces composés doivent traverser ces barrières naturelles pour être métabolisés. L'étude chez la souche *Pseudomonas fluorescens* LP6a, connue pour dégrader divers HAP comme le naphtalène ou le phénanthrène, a montré la présence de plusieurs systèmes d'imports et d'exports (Bugg *et al.*, 2000). Cette étude a ainsi mis en évidence, par des mesures de radioactivité et l'utilisation de



<u>Figure 8 :</u> Représentation schématique des mécanismes d'adaptations cellulaires contre les effets toxiques de composés organiques.

(**A**) Changements de structure au sein de la membrane cytoplasmique. (**B**) Modifications des protéines (composition et/ou concentration) au sein de la membrane. (**C**) Réduction de l'hydrophobicité cellulaire. (**D**) Exportation active des molécules organiques. (**E**) Modification du composé organique (adaptée de Isken et de Bont, 1998).

HAP marqués au C_{14}, deux systèmes de transport : un système passif du naphtalène et du phénanthrène pour entrer dans la cellule, et un système actif conçu pour rejeter le phénanthrène spécifiquement (aucune identification de transporteurs ou de facilitateurs n'ayant été réalisée dans cette étude). Ces deux systèmes fonctionnent donc en opposition l'un par rapport à l'autre pour le phénanthrène. De plus, ils ont pu démontrer que le système de transport actif est codé par des gènes situés sur le chromosome. L'expression de celui-ci est activée par la présence de phénanthrène, permettant de maintenir un équilibre au sein de la cellule. Une seconde étude sur cette souche *Pseudomonas fluorescens* LP6a a permis d'identifier les gènes codant pour le système actif de transport de la superfamille RND, impliquée dans la résistance à plusieurs antibiotiques (Hearn *et al.*, 2003). Une autre étude plus récente chez la souche *Arthrobacter* sp. Sphe3, capable de dégrader le phénanthrène, montre que ce HAP pénètre la cellule via deux mécanismes : un mécanisme de diffusion passive en présence de glucose, et un système de transport actif quand le phénanthrène est la seule source de carbone présente. Cette étude met en évidence la sélectivité de la source de carbone la plus simple à utiliser pour le microorganisme, via la régulation du système de transport actif en fonction des sources carbonées disponibles (Kallimanis *et al.*, 2007).

Cependant, l'accumulation de composés lipophiles (comme les HAP, qui sont des molécules toxiques) dans cette membrane cytoplasmique peut entraîner des effets déstructurants, que les bactéries compensent par différentes modifications, comme par exemple celles portant sur l'hydrophobicité cellulaire et sur les protéines de membrane (composition ou concentration) (Isken et de Bont, 1998; Kallimanis *et al.*, 2007) (Figure 8). Il a ainsi été montré chez les souches du genre *Pseudomonas* la synthèse d'acides gras insaturés de configuration *trans* lorsqu'elles sont exposées à ces composés. Cette isomérisation a pour effet de structurer et de diminuer la fluidité de la membrane cytoplasmique, compensant l'effet perturbateur du composé lipophile (Bugg *et al.*, 2000). Dans le cas de cette étude, cela n'empêche cependant pas la pénétration intracellulaire de la molécule, mais atténue l'effet perturbateur engendré.

5. La diversité des voies métaboliques aérobies de dégradation des HAP

Suite à la pénétration des HAP au sein des cellules des microorganismes, la machinerie enzymatique dédiée à la dégradation de ces composés peut s'enclencher. Il existe ainsi une grande diversité de voies métaboliques aérobies des HAP. Cependant, l'attaque

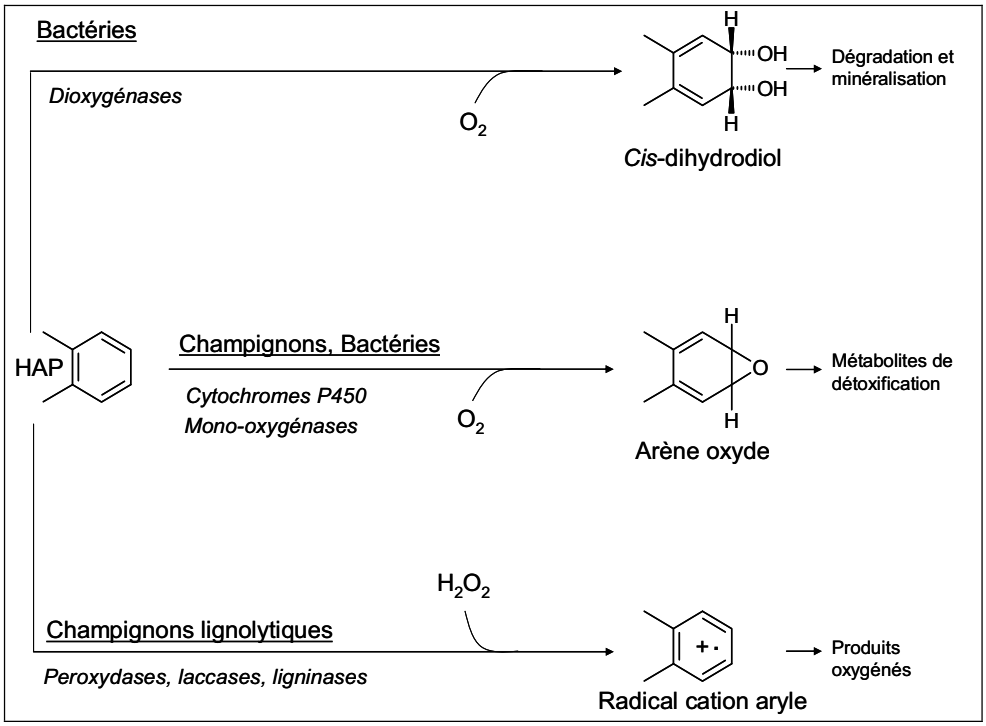


Figure 9 : Principales étapes initiales des mécanismes de dégradation par les microorganismes des hydrocarbures aromatiques polycycliques (adaptée de Haristah et Kaushik, 2009).

initiale est un indicateur précis de la voie métabolique mise en jeu. (Cerniglia, 1992; Vandecasteele, 2005). Les principaux types d'attaque, au nombre de trois, impliquent soit des dioxygénases, soit des mono-oxygénases, soit des ligninases (Figure 9).

L'attaque conduisant à la formation d'un *cis*-dihydrodiol est caractéristique des bactéries. Elle est réalisée par une dioxygénase (appelée généralement dans la littérature « ring-hydroxylating dioxygenase ») qui est une enzyme multimérique. Elle peut être composée d'une dioxygénase initiale formée d'une grande et d'une petite sous-unité (respectivement α et β), d'une sous-unité de ferrédoxine, et d'une ferrédoxine réductase dépendante du NAD(P)H, H^+ (Ferraro *et al.*, 2005; Kweon *et al.*, 2008). Elle va permettre l'incorporation de deux atomes d'oxygène au sein du noyau aromatique pour former le *cis*-dihydrodiol (Cerniglia, 1992; Vandecasteele, 2005; Haritash et Kaushik, 2009). Cette réaction nécessite de l'énergie, fournie par une ferrédoxine réductase et transmise via la sous-unité de ferrédoxine (Cerniglia, 1992; Harayama *et al.*, 1992). Enfin, c'est généralement cette enzyme, dont la spécificité est plus ou moins importante, qui va déterminer le spectre de composés qui seront dégradés par l'espèce considérée.

Deux voies d'attaques ont été décrites pour les champignons, l'une s'effectue par des mono-oxygénases qui forment des époxydes (souvent des cytochromes P450), l'autre par des enzymes lignolytiques (généralement extracellulaires). Ces différents mécanismes d'attaque, qui ne conduisent pas à une minéralisation du HAP, ont fait l'objet de plusieurs synthèses bibliographiques (Cerniglia et Heitkamp, 1989; Cerniglia, 1992; Sutherland, 1992). Chez les champignons non lignolytiques, l'attaque initiale des HAP, catalysée par une monooxygénase, se traduit par l'incorporation d'un atome d'oxygène dans un cycle aromatique du HAP, qui forme alors un arène oxyde. Ce mécanisme est le même que celui utilisé par les organismes supérieurs pour la détoxification des molécules aromatiques (Cerniglia, 1992; Vandecasteele, 2005).

Les champignons lignolytiques (comme *Phanerochaete chrysosporium* et *Pleurotus ostreatus*) possèdent quant à eux des cytochromes P450 qui leur permettent de dégrader des HAP en formant notamment des *trans*-dihydrodiols. Cependant, ils produisent également, dans certaines conditions de culture, des enzymes extracellulaires : les lignines peroxydases (ou ligninases), et les peroxydases manganèse-dépendantes. Ces enzymes ont comme rôle principal la dégradation de la matière organique, et plus particulièrement de la lignine. Ces enzymes, également très peu spécifiques, permettent, outre l'oxydation de la lignine, celle des HAP et d'autres composés organiques comme les phénols. Ces oxydations, y compris celles de la lignine, sont pour la plupart des réactions de cométabolisme (Cerniglia, 1992).

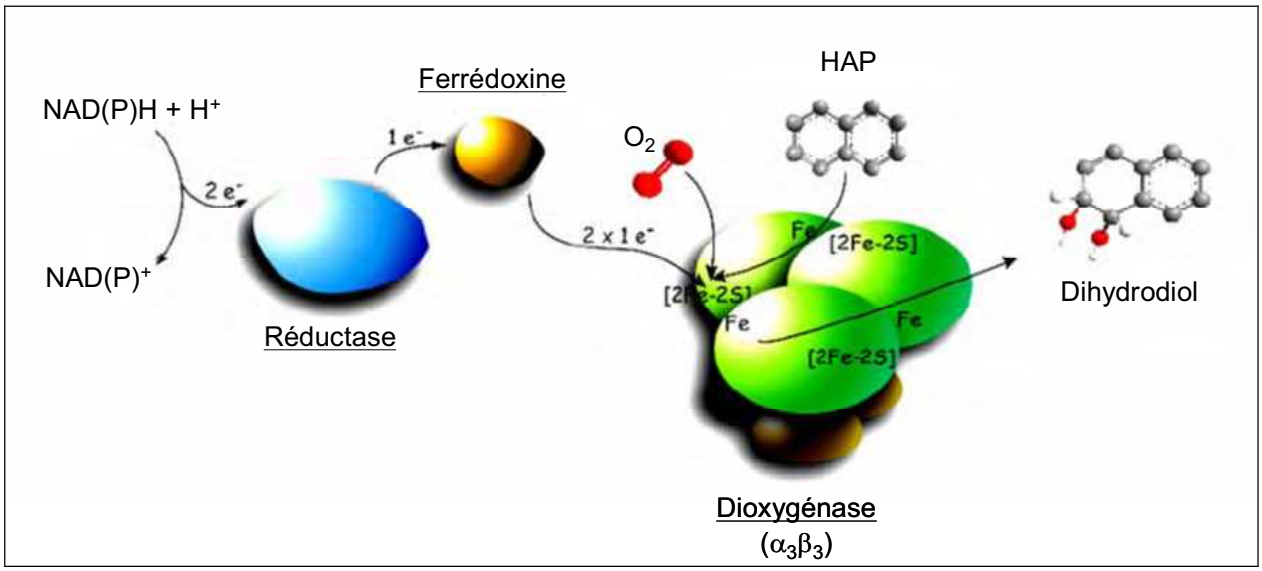


Figure 10 : Réaction catalysée par la dioxygénase initiale, enzyme multimérique impliquée dans l'attaque initiale des HAP. Les électrons provenant de l'oxydation du NAD(P)H, H^+ sont utilisés pour activer l'oxygène, par le transfert à la ferrédoxine, et l'action de la ferrédoxine réductase. Grâce à l'action de l'enzyme, le produit aromatique est alors attaqué et permet la formation d'un dihydrodiol (adaptée de Ferraro *et al*, 2005).

La dégradation par les microorganismes photosynthétiques implique des dioxygénases et des mono-oxygénases, mais ces communautés sont encore peu étudiées. Cependant, il est indispensable de noter que les microorganismes photosynthétiques (algues vertes et cyanobactéries) peuvent produire à la fois des *cis*- et des *trans*-dihydrodiols durant la dégradation des HAP (Cerniglia, 1992).

Seule la dégradation bactérienne aérobie des HAP sera abordée en détail par la suite, car la majorité des dégradations bactériennes ne sont pas des réactions de co-métabolisme, ces dernières étant plus difficiles à mettre en évidence, comme c'est le cas pour les champignons ou les microorganismes photosynthétiques.

6. Les dioxygénases, des enzymes clés de la dégradation des HAP

La dégradation aérobie des HAP par les bactéries commence généralement par l'incorporation de deux atomes d'oxygène au sein d'un cycle aromatique, pour produire un dihydrodiol. Cette étape, qui permet le clivage du cycle, est souvent la plus difficile, car le composé à dégrader est formé de cycles aromatiques non substitués très stables (dans le cas des HAP). Cette attaque initiale est généralement catalysée par des dioxygénases qui sont des enzymes multimériques pour les microorganismes bactériens aérobies (Pinyakong *et al.*, 2003a; Ferraro *et al.*, 2005; Kweon *et al.*, 2008; Baek *et al.*, 2009).

6.1. Mécanisme d'action moléculaire des (di-)oxygénases

Ces dioxygénases sont formées par trois composants : une ferrédoxine, une ferrédoxine réductase dépendante du NAD(P)H, H^+ et une dioxygénase, généralement composée de deux sous-unités d'oxygénase : une grande (nommée α) et une petite (nommée β) (Ferraro *et al.*, 2005) (Figure 10). Le plus souvent, trois grandes sous-unités d'oxygénase sont associées (α_3), ces dernières pouvant également être associées avec trois petites sous-unités d'oxygénase ($\alpha_3\beta_3$). Les structures des dioxygénases peuvent donc être de type α_3 ou $\alpha_3\beta_3$ (Ferraro *et al.*, 2005). La sous-unité α contient deux régions conservées (qui sont les deux centres redox) : un cluster [2Fe-2S] (regroupant deux ions fer et deux ions soufre) et un ion ferreux seul (adjacent à la poche catalytique) (Jakoncic *et al.*, 2007a, b). Ces sous-unités α sont connues pour être les composants catalytiques impliquées dans le transfert des électrons aux molécules d'oxygène (Kweon *et al.*, 2008). De plus, c'est de cette protéine que va dépendre le spectre de substrat de la dioxygénase car elle exerce une contrainte sur la position du substrat au sein de la poche catalytique (Ferraro *et al.*, 2005; Jakoncic *et al.*, 2007a, b). La

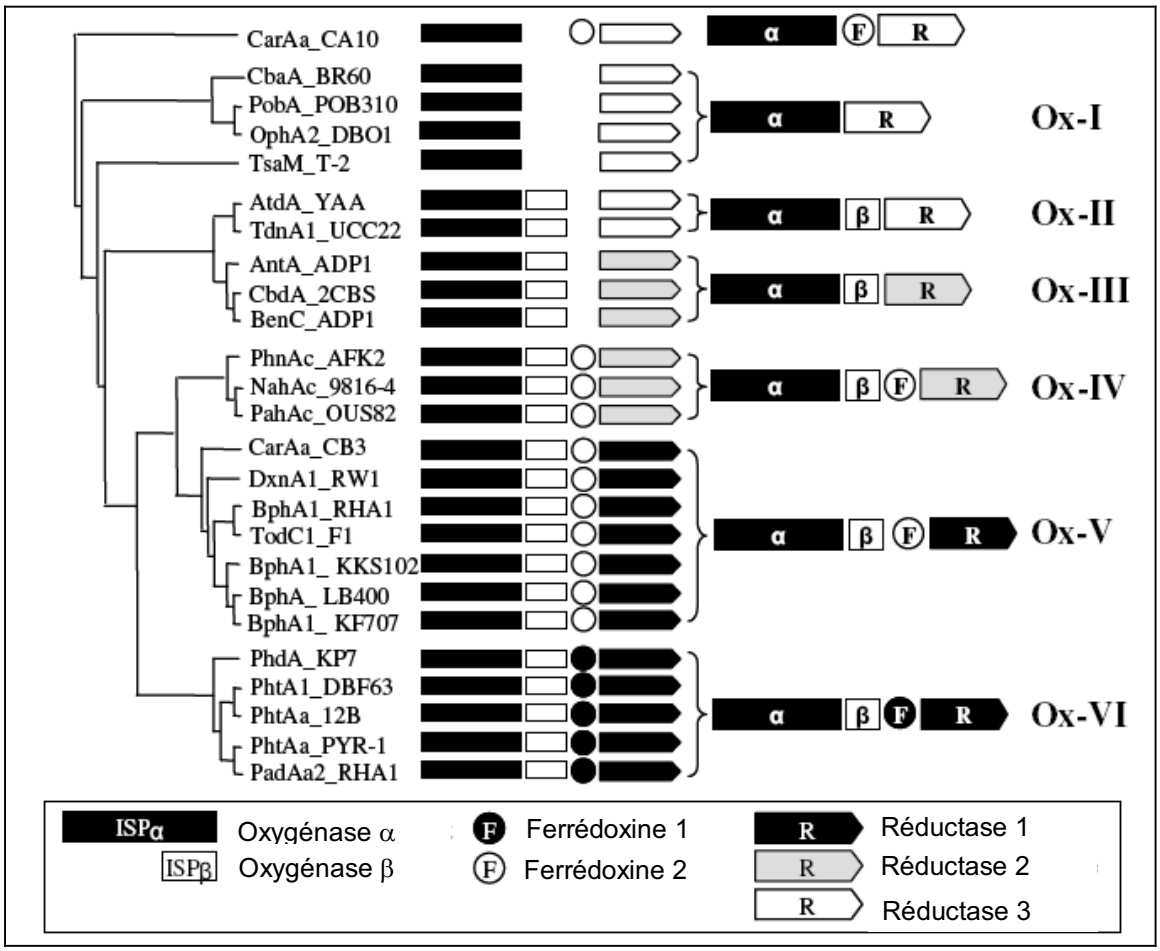
sous-unité β, quant à elle, est supposée servir de stabiliseur à la sous-unité α, mais pourrait également jouer un rôle dans la reconnaissance au substrat (Kumar et Khanna, 2010).

Les électrons provenant de l'oxydation du NAD(P)H, H^+ sont utilisés pour activer l'oxygène, permettant l'oxydation du substrat (Figure 10). C'est pourquoi, deux autres protéines sont indispensables : une ferrédoxine et une ferrédoxine réductase dépendante du NAD(P)H, H^+. La réductase oxyde le NAD(P)H, H^+ pour former le $NAD(P)^+$, capturant à travers cette réaction deux électrons. Chaque électron va permettre la réduction de la ferrédoxine, qui va ensuite le transférer à l'oxygénase (cette réaction a lieu deux fois pour transférer les deux électrons) (Ferraro *et al.*, 2005). Ces enzymes sont classées en fonction de la composition atomique du cluster [Fe-S] présent, les plus souvent retrouvées étant les clusters [2Fe-2S] et [3Fe-4S] (Saito *et al.*, 2000; Kim *et al.*, 2006; Jakoncic *et al.*, 2007a, b).

6.2. Classifications des dioxygénases

Les dioxygénases initiales étant des enzymes de grande importance pour la dégradation des composés aromatiques, elles ont été fortement étudiées, leurs spectres de substrats évalués, et sont même utilisées comme marqueurs fonctionnels (Hernandez-Raquet *et al.*, 2006; Kim et Crowley, 2007b; Cébron *et al.*, 2008; Lozada *et al.*, 2008; Kumar et Khanna, 2010). La plupart des dioxygénases initiales, bien que montrant une certaine similarité de séquence, ont un spectre de substrat qui varie énormément (Nam *et al.*, 2001). Depuis 1992, diverses approches ont tenté de regrouper ces dioxygénases en classes. Ainsi, une première classification basée sur la nature des centres redox et des composants enzymatiques a donné lieu à la formation de trois classes différentes, mais difficile à considérer comme une classification phylogénétique à proprement parler (Batie *et al.*, 1992).

Avec la découverte croissante de nouvelles séquences codant pour ces enzymes, une nouvelle classification a été nécessaire. Cette dernière, basée sur le pourcentage de similarité des séquences protéiques des sous-unités α a permis la formation de quatre classes différentes, qui regroupent ces protéines. La classe I regroupe des séquences avec un large spectre de substrats, la classe II regroupe les dioxygénases spécifiques du benzoate et du toluate, la classe III regroupant celles attaquant les HAP comme le naphtalène ou le phénanthrène et la classe IV spécifique du toluène, du benzène et du biphényle (Nam *et al.*, 2001). Enfin, une nouvelle classification, proposée récemment par Kweon et ses collaborateurs (Kweon *et al.*, 2008), regroupe ces deux systèmes de classification en décidant d'analyser la dioxygénase comme un ensemble. En effet, cette nouvelle classification est basée à la fois sur l'analyse des différentes sous-unités de l'enzyme (ferrédoxine réductases,



<u>Figure 11 :</u> Représentation graphique de l'organisation en plusieurs classes oxydatives (de I à VI) des dioxygénases. (Proposé par Kweon *et al.,* 2008).

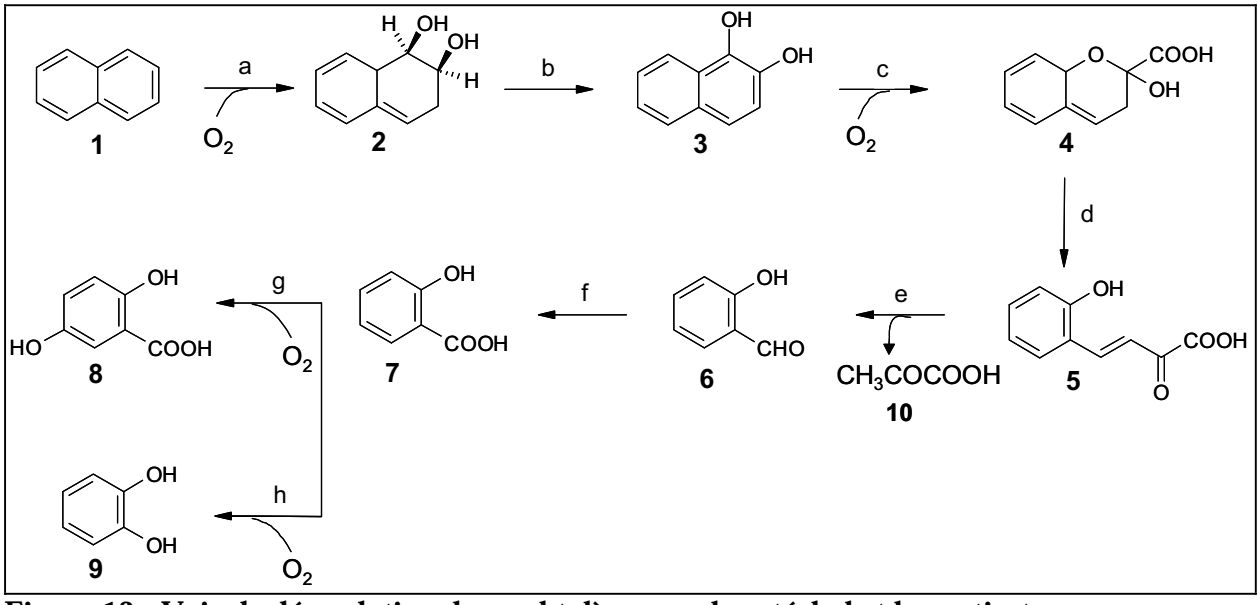
Séquences : (CarAa_CA10) Carbazole 1,9a-dioxygénase de *P. resinovorans* CA10 (D89064) ; (CbaA_BR60) et (PobA_POB310) Phénoxybenzoate dioxygénase d'*Alcaligenes* sp. BR60 (U18133) et de *P. pseudoalcaligenes* POB310 (X78823) ; (OphA2_DB01) Phthalate dioxygénase de *Burkholderia cepacoa* DBO1 (AF095748) ; (TsaM_T-2) *p*-Toluènesulfonate monooxygénase de *Comamonas testosteroni* T-2 (U32622) ; (AtdA_YAA) Aniline dioxygénase de *Acinetobacter* sp. YAA (D86080) ; (TdnA1_UCC22) Aniline oxygénase de *P. putida* UCC22 (D85415) ; (AntA_ADP1) Anthranilate dioxygénase de *Acinetobacter*. sp. ADP1 (AF071556) ; (CbdA_2CBS) 2-Halobenzoate 1,2-dioxygénase de *P. cepacia* 2CBS (X79076) ; (BenC_ADP1) Benzoate 1,2-dioxygénase de *Acinetobacter*. sp. ADP1 (AF009224) ; (PhnAc_AFK2) 3,4-Dihydroxyphénanthrène dioxygénase d' *A. faecalis* AFK2 (AB024945) ; (NahAc_9816-4) Naphthalène dioxygénase de *Pseudomonas* sp. 9816-4 (U49496) ; (PahAc_OUS82) PAH dioxygénase de *P. putida* OUS82 (AB004059) ; (CarAa_CB3) Carbazole dioxygénase de *Sphingomonas* sp. CB3 (AF060489) ; (DxnA1_RW1) Dioxin dioxygénase de *Sphingomonas* sp. RW1 (X72850) ; (BphA1_RHA1) Biphényle dioxygénase de *Rhodococcus* sp. RHA1 (D32142) ; (TodC1_F1) Toluène dioxygénase de *P. putida* F1 (J04996) ; (BphA1_KKS102) et (BphA1_KF707) Biphényle dioxygénase de *Pseudomonas* sp. KKS102 (D17319) et de *P. pseudoalcaligenes* KF707 (M83673) ; (BphA_LB400) Biphényle 2,3-dioxygénase de *B. xenovorans* LB400 (M86348) ; (PhdA_KP7) Phénanthrène dioxygénase de *Nocardioides* sp. KP7 (AB017795) ; (PhtA1_DBF63), (PhtAa_12B), (PhtAa_PYR-1) et (PadAa2_RHA1) Phthalate dioxygénase de *Terrabacter* sp. DBF63 (AB084235), de *A. keyseri* 12B (AF331043), de *M. vanbaalenii* PYR-1 (AY365117) et de *Rhodococcus* sp. RHA1 (AB154537).

ferrédoxines, etc.…) lorsqu'elles sont connues et identifiées, ainsi que sur les séquences protéiques disponibles des sous-unités α. En se basant sur ces deux critères, il est possible de regrouper les dioxygénases en six groupes distincts (Figure 11). C'est cette classification qui semble permettre d'avoir à la fois une vision de conservation de séquence (en se basant sur les similarités de séquences des sous-unités α) mais aussi plus enzymatique, en considérant les différentes sous-unités impliquées (domaines conservés, composition, identité de séquences, etc.…). Cette classification permet donc d'organiser les dioxygénases initiales en fonction de leur composition, mais également de leur spectre de substrat et de leurs homologies.

Cette approche a permis la classification de 130 dioxygénases, en se basant sur 25 d'entre elles bien connues et considérées comme des « standards ». Ces données ont servi de bases pour l'implémentation d'un programme informatique permettant une classification automatique des nouvelles enzymes identifiées et une évolution de cette classification avec l'ajout de nouvelles séquences et de nouvelles informations par l'utilisateur (Baek *et al.*, 2009). Cet outil se base cependant uniquement sur les similarités entre les séquences de sous-unités α déjà classées et celles données par l'utilisateur.

7. La dégradation bactérienne des HAP

Les voies de dégradation bactériennes des HAP sont composées pour la plupart de deux phases bien définies : une phase spécifique, qui permet l'attaque initiale du HAP considéré, entraînant la formation de métabolites comme le catéchol, ou un de ses précurseurs (comme le protocatéchuate, ou l'acide salicylique) (Seo *et al.*, 2009). La seconde phase de dégradation est commune à plusieurs composés aromatiques (comme les BTEX par exemple), permettant ainsi de réduire le nombre de gènes nécessaires à la dégradation de différents HAP. Cette phase est généralement appelée la « voie basse ». C'est cette dernière qui permet de produire les substrats du métabolisme cellulaire central. Elle peut être réalisée par deux voies distinctes, engagées par une fission du cycle par des dioxygénases, l'une dite *meta* et l'autre *ortho*, qui vont être détaillées. Chez certains microorganismes, la dégradation ne passe pas par la formation de catéchol mais par celle du gentisate (proche du catéchol), dont le précurseur est l'acide salicylique. Ce gentisate peut également permettre la production d'énergie via une voie annexe particulière (Zhou *et al.*, 2001; Jeon *et al.*, 2006).



<u>Figure 12 :</u> Voie de dégradation du naphtalène vers le catéchol et le gentisate.

Composés : (**1**) Naphtalène ; (**2**) *cis*-1,2-naphtalène dihydrodiol ; (**3**) 1,2-dihydroxynaphtalène ; (**4**) 2-hydroxychromène-2-carboxylate; (**5**) *trans-o*-hydroxybenzylidène pyruvate ; (**6**) salicylaldéhyde ; (**7**) salicylate ; (**8**) gentisate ; (**9**) catéchol ; (**10**) pyruvate.

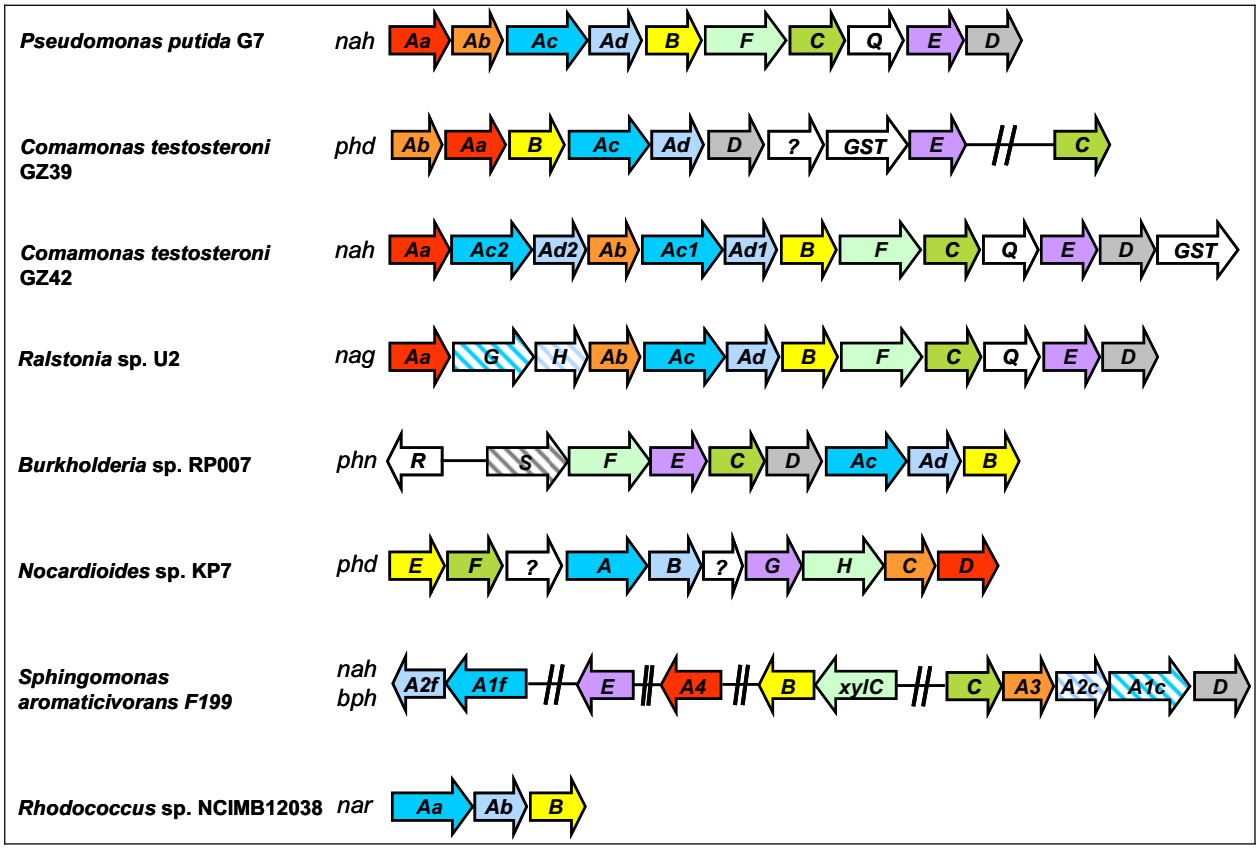
Enzymes et nomenclature des gènes correspondants donnée chez *Pseudomonas putida* G7 : (a) Naphtalène-1,2-dioxygénase (*nahAaAbAcAd*) ; (b) *cis*-1,2-naphtalène dihydrodiol déshydrogénase (*nahB*) ; (c) 1,2-dihydroxynaphtalène dioxygénase (*nahC*) ; (d) 2-hydroxychromène-2-carboxylate isomérase (*nahD*), réaction réversible ; (e) *trans-o*-hydroxybenzylidène pyruvate hydratase-aldolase (*nahE*) ; (f) salicylaldéhyde déshydrogénase NAD-dépendante (*nahF*) ; (g) salicylate-5-hydroxylase ; (h) salicylate hydroxylase (*nahG*) (adaptée de Seo *et al*, 2009).

7.1. Voie « haute » de dégradation des HAP à deux cycles : exemple du naphtalène

Le métabolisme du naphtalène par les *Pseudomonas* a été décrit pour la première fois en 1964 (Davies et Evans, 1964). A l'heure actuelle, la dégradation bactérienne aérobie du naphtalène est très bien caractérisée (Goyal et Zylstra, 1997). Les voies cataboliques mises en jeu ont été d'abord étudiées chez *Pseudomonas putida* PpG7 qui contient le plasmide de dégradation NAH7, et chez plusieurs autres espèces de *Pseudomonas* qui contiennent des plasmides portant le même opéron (Habe et Omori, 2003). De nombreuses souches d'autres genres (comme *Burkholderia, Sphingomonas, Mycobacterium, Polaromonas, Ralstonia* et *Rhodococcus*) ont ensuite été isolées et étudiées pour leur capacité à dégrader ce substrat (Tableau 3) (Kelley *et al.*, 1990; Larkin *et al.*, 1999; Laurie et Lloyd-Jones, 1999; Kulakov *et al.*, 2000; Zhou *et al.*, 2001).

L'étape initiale de la dégradation du naphtalène (Figure 12) est réalisée par une dioxygénase initiale pour former le *cis*-1,2-naphtalène dihydrodiol (Cerniglia, 1992; Haritash et Kaushik, 2009). Le métabolite obtenu est ensuite oxydé par une *cis*-1,2-dihydrodiol déshydrogénase en 1,2-dihydroxynaphtalène (Habe et Omori, 2003; Peng *et al.*, 2008; Seo *et al.*, 2009). Une nouvelle réaction d'oxygénation dite *meta* ouvre alors le cycle aromatique dihydroxylé formant l'acide 2-hydroxy-2H-chromène-2-carboxylique. Deux enzymes (une isomérase et une hydratase-aldolase) permettent la formation de salicylaldéhyde qui sera transformé en acide salicylique via une salicylaldéhyde déshydrogénase. Enfin, l'acide salicylique ainsi formé peut, soit être converti en catéchol, soit en gentisate, pour intégrer la voie « basse » (voir paragraphe **La production d'énergie via : les voies *ortho*, *meta* ou du gentisate** page 39). Les bactéries réalisant cette voie haute de dégradation du naphtalène sont largement distribuées au sein des écosystèmes. Les voies métaboliques, les enzymes impliquées et les régulations mises en jeu ont ainsi été étudiés en détail (Cerniglia, 1992; Habe et Omori, 2003; Peng *et al.*, 2008; Seo *et al.*, 2009). Pour la souche *Pseudomonas putida* G7, deux opérons existent au sein du plasmide NAH7. Le premier, l'opéron *nah* codant la voie « haute », permet de transformer le naphtalène en salicylate et est organisé au sein d'un seul opéron. Le second opéron correspond à la voie « basse », décrite plus loin (voir paragraphe **La production d'énergie via : les voies *ortho*, *meta* ou du gentisate** page 39). La régulation de ces deux opérons est contrôlée par la protéine NahR, située entre les deux opérons.

Chez les autres genres étudiés, certaines divergences ont été caractérisées, portant notamment sur l'organisation génomique des gènes (réorganisation(s), délétion(s) ou



<u>Figure 13 :</u> Représentation de l'organisation génétique des opérons portant les gènes codant pour les protéines impliquées dans la voie « haute » de dégradation du naphtalène de différents organismes.

Le nom de chaque gène est indiqué (en se basant sur la définition de chaque auteur). La taille des flèches ne représente pas la taille réelle de chacun des gènes. Ces données sont basées sur les informations de séquences disponibles pour chaque espèce : *Pseudomonas putida* G7 (NC_007926), *Comamonas testosteroni* GZ39 et *Comamonas testoteroni* GZ42 (Goyal et Zylstra, 1997), *Ralstonia* sp. U2 (AF036940), *Burkholderia* sp. RP007 (AF061751), *Nocardioides* sp. KP7 (AB031319), *Sphingomonas aromaticivorans* F199 (NC_002033) et *Rhodococcus* sp. NCIMB12038 (AF082663). Un code couleur a été utilisé pour différencier l'activité des protéines codées par les gènes indiqués. Rouge : ferrédoxine réductase, orange : sous-unité de ferrédoxine, bleu : grande et petite sous-unité de la dioxygénase initiale, respectivement α et β, bleu hachuré : grande et petite sous-unité d'oxygénase, jaune : dihydrodiol déshydrogénase, vert clair : déshydrogénase, vert foncé : extradiol dioxygénase, violet : hydratase-aldolase, gris : isomérase, hachuré en gris : régulateurs potentiels, blanc : inconnu.

insertion(s)), mais également sur les séquences elles-mêmes, qui peuvent présenter de grandes variabilités, traduisant un fort éloignement phylogénétique (Figure 13). C'est notamment le cas chez les souches de *Comamonas testosteroni* (comme GZ38A, GZ39 et GZ42), ayant la capacité de dégrader le naphtalène, et dont les opérons ont été séquencés (Goyal et Zylstra, 1997). Les gènes de la voie haute chez la souche GZ39 (cluster *phd*) sont organisés différemment de ceux chez *Pseudomonas putida* G7 (Figure 13) (Goyal et Zylstra, 1997; Habe et Omori, 2003; Peng *et al.*, 2008). On note par exemple la présence de *phdB* au sein du groupe *phdAbAaAcAd*, l'inversion de *phdAb* et *phdAa*, mais aussi l'absence du gène codant la 1,2-dihydroxynaphtalène dioxygénase au sein de l'opéron (le gène étant présent ailleurs sur le plasmide), et enfin la présence du gène codant une glutathione-S-transférase (Goyal et Zylstra, 1997; Peng *et al.*, 2008). Au contraire, pour une autre souche de *Comamonas*, la souche GZ42, bien que les gènes retrouvés chez *Pseudomonas putida* G7 soient tous présents et présentent pratiquement la même organisation génomique, les séquences sont relativement divergentes (Figure 13) (Goyal et Zylstra, 1997). Il faut noter la présence de deux autres gènes uniquement chez la souche GZ42, *nahAc2* et *nahAd2* (codant les sous-unités α et β d'une autres dioxygénase potentielle), situés entre *nahAa* et *nahAb*. Ils sont suspectés de ne pas être fonctionnels, bien que cela n'ait pas encore été vérifié (Goyal et Zylstra, 1997; Habe et Omori, 2003).

Les séquences de ces deux gènes *nahAc2* et *nahAd2* présentent de fortes similarités avec les gènes *nagG* et *nahH* (codant deux sous-unités d'oxygénase d'une salicylate-5-hydroxylase), isolés chez *Rasltonia* sp. U2 et convertissant le salicylate en gentisate (Figure 13) (Zhou *et al.*, 2001). Cette organisation est similaire à celle connue pour l'opéron *nah* de *Pseudomonas putida*, hormis l'insertion des deux gènes *nagG* et *nagH*.

Un nouvel opéron a aussi été caractérisé chez *Burkholderia* sp. RP007 (Laurie et Lloyd-Jones, 1999; Habe et Omori, 2003) (Figure 13). On note tout d'abord la présence de deux gènes codant pour des protéines régulatrices : *phnS*, en début d'opéron, et en amont et situé sur l'autre brin par rapport à cet opéron, le gène *phnR*. Le mécanisme de régulation de cet opéron n'a cependant pas encore été décrit précisément. Il est intéressant de noter l'absence de gènes codant une ferrédoxine et une ferrédoxine réductase au sein de cet opéron (Figure 13). Il ne peut être exclu que ces gènes aient une autre localisation, sachant que le génome de cette souche n'a actuellement pas été séquencé.

Au sein du genre *Mycobacterium*, de nombreuses souches sont également capables de dégrader le naphtalène (Kelley *et al.*, 1990; Kim *et al.*, 2007), et bien que la voie métabolique utilisée semble être la même (Kelley *et al.*, 1990), les gènes sont localisés au sein de plusieurs

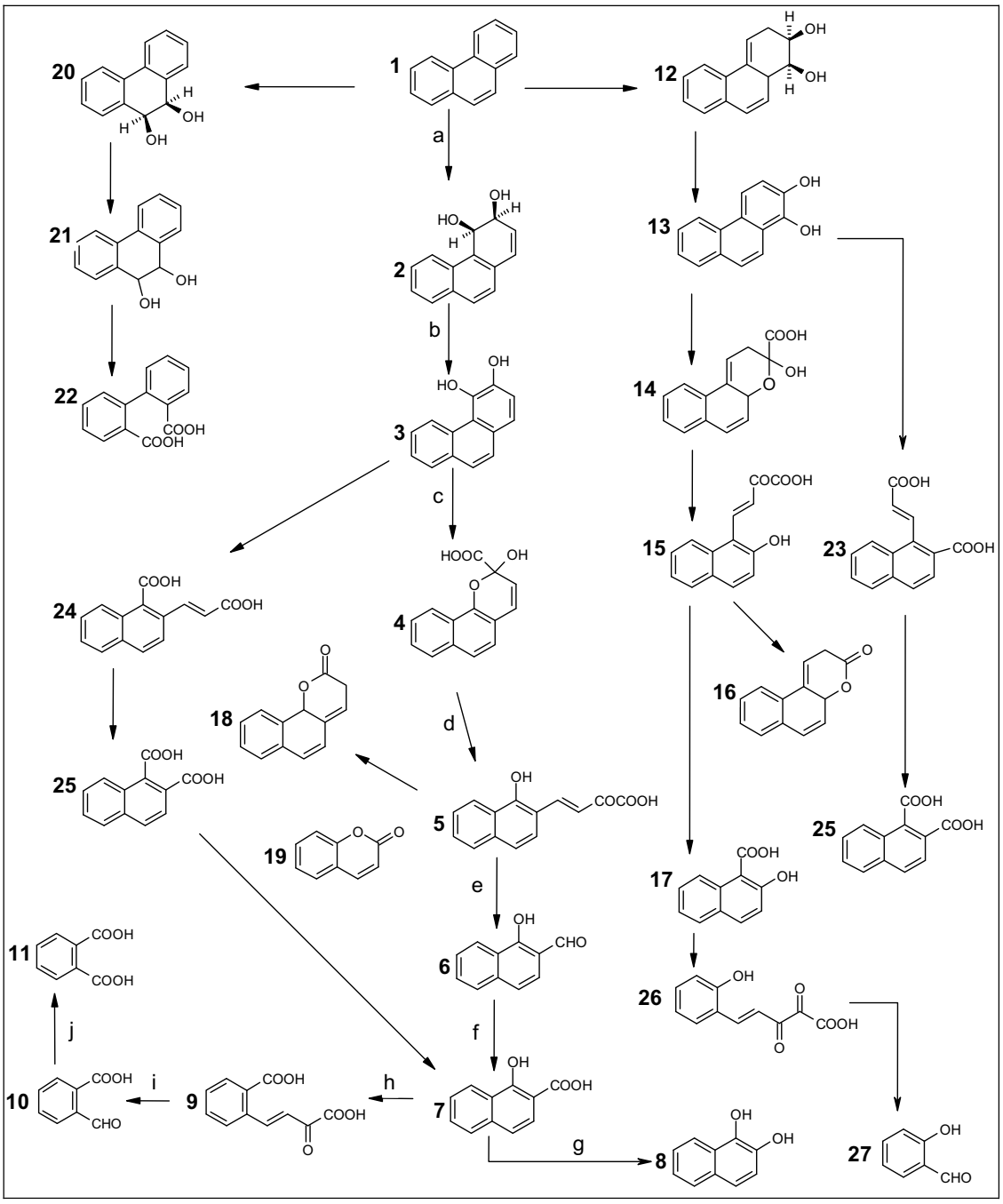


Figure 14 : Voies de dégradations potentielles du phénanthrène.

Les flèches sans lettres peuvent représenter une ou plusieurs étapes enzymatiques inconnues, seuls les métabolites représentés ayant été détectés.

Composés : (**1**) Phénanthrène ; (**2**) *cis*-3,4-phénanthrène dihydrodiol ; (**3**) 3,4-dihydroxyphénanthrène ; (**4**) 2-hydroxybenzo(*h*)chromène-2-carboxylate ; (**5**) 4-(1-hydroxynapht-2-yl)-2-oxobut-3-énoate ; (**6**) 1-hydroxy-2-naphtaldéhyde ; (**7**) 1-hydroxy-2-napthoate ; (**8**) 1,2-dihydroxynaphtalène ; (**9**) *trans*-*o*-carboxybenzylidène pyruvate ; (**10**) 2-carboxybenzaldéhyde ; (**11**) *o*-phthalate ; (**12**) *cis*-1,2-phénanthrène dihydrodiol ; (**13**) 1,2-dihydroxyphénanthrène ; (**14**) 3-hydroxy-3H-benzo(*f*)chromène-3-carboxylate ; (**15**) *cis*-4-(2-hydroxynaph-1-yl)-2-oxobut-3-énoate ; (**16**) 5,6-benzocoumarine ; (**17**) acide 2-hydroxy-1-naphtoïque ; (**18**) 7,8-benzocoumarine (**19**) benzocoumarine ;(**20**) cis-9,10-phénanthrène dihydrodiol ; (**21**) 9,10-dihydroxyphénanthrène ; (**22**) acide 2,2'-diphénique ; (**23**) 1-(2-carboxy-vinyl)-naphtalène-2-carboxylate ; (**24**) 2-(2-carboxy-vinyl)-naphtalène-1-

opérons (Kim *et al.*, 2007). Enfin, chez la souche *Nocardioides* sp. KP7 (ce genre étant le plus étudié pour les bactéries Gram-positive en ce qui concerne la dégradation des HAP), une organisation génétique a été proposée, totalement différente de celle de *Pseudomonas putida* G7 (Figure 13) (Saito *et al.*, 1999) . De plus, les gènes composant cet opéron, bien que reliés à ceux connus pour le genre *Pseudomonas*, ont tout de même une faible identité de séquence (moins de 60 % pour certaines enzymes) (Saito *et al.*, 1999; Habe et Omori, 2003).

Une autre organisation génétique a été mise en évidence chez d'autres souches, comme *Sphingomonas yanoikuyae* B1 et *Sphingomonas aromaticivorans* F199 (Zylstra et Kim, 1997; Romine *et al.*, 1999b) (Figure 13). Ces souches, comme de nombreuses autres du genre *Sphingomonas*, sont capables d'utiliser de nombreux substrats aromatiques (Pinyakong *et al.*, 2003a; Stolz, 2009). Le séquençage complet du plasmide de 184 kpb de la souche F199 a permis de caractériser 15 clusters regroupant des gènes de dégradation de différents composés aromatiques (Romine *et al.*, 1999b; Pinyakong *et al.*, 2003a). Cette disposition permet vraisemblablement de mettre à profit des mécanismes biochimiques similaires (réalisés par diverses isoenzymes codées par des gènes situés sur plusieurs opérons), pour dégrader différents hydrocarbures aromatiques en mélange. Par exemple, certains gènes de la voie de dégradation du naphtalène et du *m*-xylène sont nécessaires pour la dégradation du biphényle. De plus, le fait de retrouver sept couples de sous-unité d'oxygénases ou de dioxygénases au sein de ce plasmide, confère probablement à la souche une grande capacité d'adaptation aux conditions oligotrophes spécifiques des environnements où ces souches ont été isolées (Stolz, 2009).

Enfin, pour le genre *Rhodococcus*, seuls quelques gènes ont pu être isolés chez la souche NCIMB12038 (Figure 13). Il s'agit des gènes *narAa, narAb* et *narB* codant respectivement les sous-unités α et β d'une dioxygénase initiale et d'une dihydrodiol déshydrogénase (Larkin *et al.*, 1999; Kulakov *et al.*, 2000). L'expression de ces gènes, localisés au sein de la même unité transcriptionnelle, est dépendante du naphtalène (Kulakov *et al.*, 2000; Peng *et al.*, 2008).

7.2. Les HAP à trois cycles : exemple du phénanthrène

Plusieurs voies de dégradation bactériennes aérobies du phénanthrène ont été mises en évidence (Rothmel *et al.*, 1990; Goyal et Zylstra, 1997; Pinyakong *et al.*, 2000; Moody *et al.*, 2001; Krivobok *et al.*, 2003; Seo *et al.*, 2006; Mallick *et al.*, 2007; Seo *et al.*, 2007; Seo *et al.*, 2009; Stolz, 2009) au sein d'une grande variété de souches appartenant à de nombreux genres bactériens (Figure 14) (Tableau 3).

carboxylate ; (**25**) naphtalène-1,2-carboxylate ; (**26**) *trans*-2,3-dioxo-5-(2'-hydroxyphényle)-pent-4-énoate ; (**27**) salicylaldéhyde.

Enzymes et nomenclature des gènes correspondants donnée chez *Nocardioides* KP7 (gènes *phd*) et *Sphingomonas yanoikuyae* B1 (gènes *bph*) : (a) Phénanthrène dioxygénase (*phdABCD* correspondant respectivement aux sous-unités d'oxygénase α et β, à la ferrédoxine et à la ferrédoxine réductase ; (b) dihydrodiol déshydrogénase (*phdE*) ; (c) extradiol dioxygénase (*phdF*) ; (d) 2-hydroxybenzo(*h*)chromène-2-carboxylate isomérase (pas de gène *phd*) ; (e) une hydratase-aldolase (*phdG*) ; (f) aldéhyde déshydrogénase (*phdH*) ; (g) salicylate oxygénase (*bphA1cA2cA3A4* correspondant respectivement aux sous-unités α et β d'oxygénase, à la ferrédoxine et à la ferrédoxine réductase) ; (h) 1-hydroxy-2-napthoate dioxygénase (*phdI*) ; (i) *trans-o*-carboxybenzylidène pyruvate hydratase-aldolase (*phdJ*) ; (j) 2-carboxybenzaldéhyde déshydrogénase (*phdK*) (adaptée de Seo *et al*, 2009).

Il arrive fréquemment qu'un produit de dégradation d'un HAP à grand nombre de cycles passe par la voie de dégradation d'un HAP à plus petit nombre de cycles (ainsi, certains produits de dégradation du phénanthrène vont pouvoir être dégradés via les enzymes de dégradation du naphtalène, et donc la voie du naphtalène). De plus, pour une grande majorité de bactéries, la dégradation du phénanthrène nécessite les mêmes étapes biochimiques (et souvent les mêmes enzymes) que la dégradation du naphtalène (voir Figures 12 et 14). Une dioxygénase initiale catalyse tout d'abord une attaque majoritairement effectuée en position 3,4, entraînant la formation du *cis*-3,4-phénanthrène dihydrodiol. Ce dernier est converti 3,4-dihydroxyphénanthrène via l'action d'une déshydrogénase (Pinyakong *et al.*, 2000; Seo *et al.*, 2009). Ce composé va ensuite être clivé en position *meta* pour former un acide *trans*-4-(1-hydroxynaph-2-yl)-2-oxobut-3-enoïque via l'action d'une dioxygénase et d'une isomérase. Enfin, grâce à l'action d'une hydratase-aldolase et d'une aldéhyde déshydrogénase, l'acide 1-hydroxy-2-naphthoïque va être formé (Figure 14). Ce composé est le point de départ de deux voies : celle de Evans (encore appelée voie des *Pseudomonas* car étudiée principalement chez ces bactéries) et celle de Kiyohara (Kiyohara et Nagao, 1978).

La voie de Evans mise en jeu permet de rejoindre la voie « haute » de dégradation du naphtalène, en transformant l'acide 1-hydroxy-2-naphthoïque en 1,2-dihydroxynaphtalène à l'aide d'une salicylate oxygénase. Chez *Sphingomonas yanoikuyae* B1, cette enzyme multimérique est codée par *bphA2cA1c, bphA3* et *bphA4* (Cho *et al.*, 2005). La molécule ainsi obtenue suit ensuite la voie de dégradation décrite pour le naphtalène (Figure 12).

La voie de Kiyohara, décrite chez *Nocardioides* sp. KP7 (Iwabuchi et Harayama, 1998a, b), nécessite trois étapes enzymatiques réalisées successivement par une 1-hydroxy-2-naphtoate dioxygénase (codée par *phdI*), une hydratase/aldolase (codé par *phdJ*) et une déshydrogénase (codé par *phdK*). Ces trois étapes permettent d'aboutir au *o*-phthalate, qui sera transformé en acide protocatéchuique via une phthalate-4,5-dioxygénase, une phthalate 4,5-dihydrodiol-déshydrogénase et une 4,5-dihydroxy-phthalate-décarboxylase (Goyal et Zylstra, 1997). Les gènes permettant la formation de l'acide 1-hydroxy-2-naphthoïque semblent organisés en un seul opéron (dont les gènes sont également codent pour des protéines impliquées dans la dégradation du naphtalène) localisé 6,1 kpb en aval de celui regroupant les trois gènes *phdI, phdJ* et *phdK* (permettant la formation du *o*-phthalate à partir de l'acide 1-hydroxy-2-naphthoïque) (Figure 14).

D'autres produits de dégradation intermédiaires ont été isolés chez plusieurs organismes comme *Sphingomonas* sp. P2, *Mycobacterium* sp. PYR-1*, Arthrobacter* sp. P1-1,

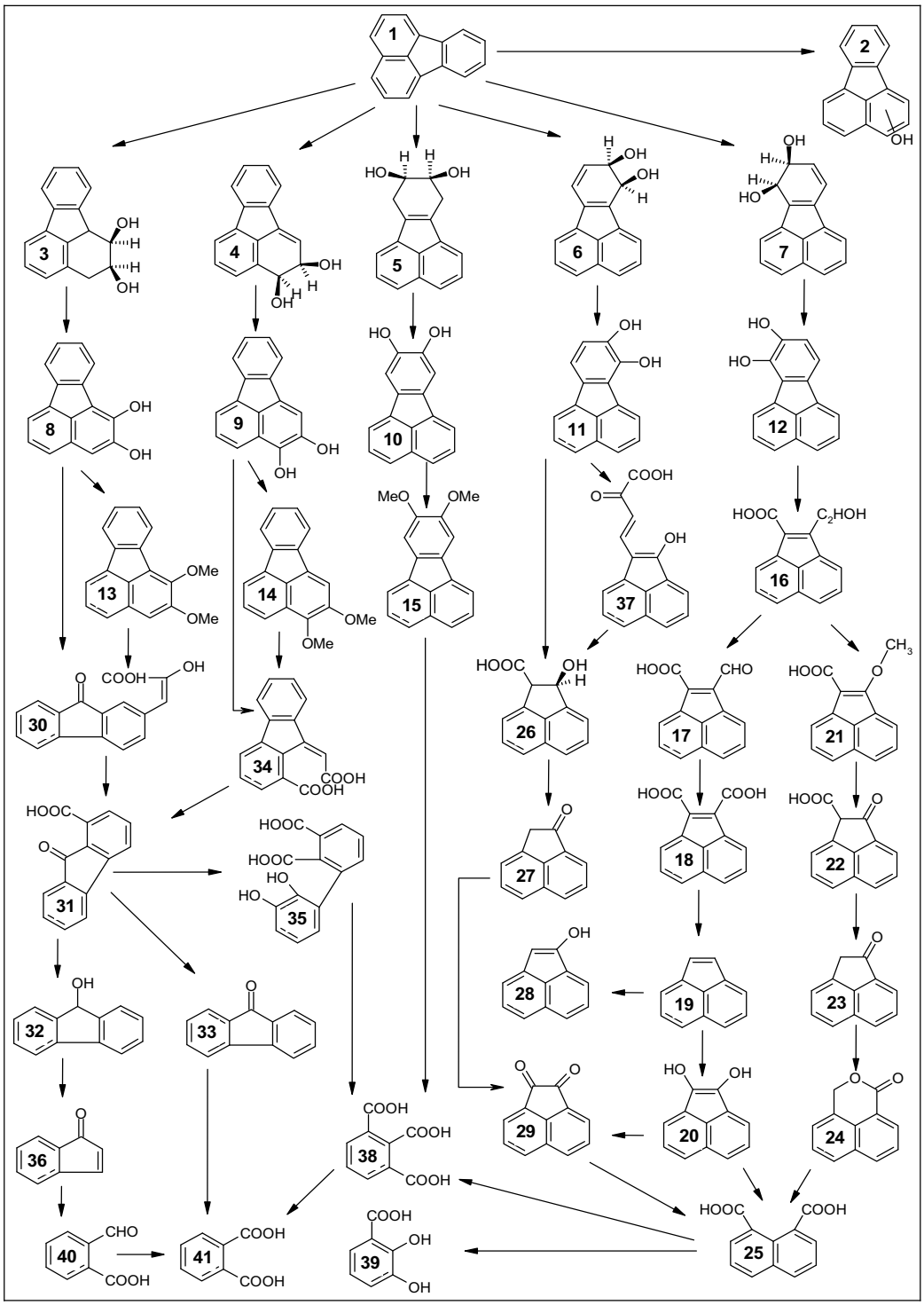


Figure 15 : Voies de dégradations potentielles du fluoranthène.

Les flèches peuvent représenter une ou plusieurs étapes enzymatiques, seuls les métabolites représentés ayant été détectés.

Composés : (**1**) Fluoranthène ; (**2**) monohydroxyfluoranthène ; (**3**) *cis*-1,2-fluoranthène dihydrodiol ; (**4**) *cis*-2,3-fluoranthène dihydrodiol ; (**5**) *cis*-9,10-fluoranthène dihydrodiol ; (**6**) *cis*-8,9-fluoranthène dihydrodiol ; (**7**) *cis*-7,8-fluoranthène dihydrodiol ; (**8**) 1,2-dihydroxyfluoranthène ; (**9**) 2,3-dihydroxyfluoranthène; (**10**) 9,10-dihydroxyfluoranthène ; (**11**) 8,9-dihydroxyfluoranthène ; (**12**) 7,8-dihydroxyfluoranthène ; (**13**) 1,2-diméthoxyfluoranthène ; (**14**) 2,3-diméthoxyfluoranthène ; (**15**) 9,10-diméthoxyfluoranthène ; (**16**) acide 2-(hydroxyméthyl)-acénapthylène-1-carboxylique ; (**17**) acide 2-

Burkholderia sp. C3, *Staphylococcus* sp. PN/Y (Pinyakong *et al.*, 2000; Moody *et al.*, 2001; Seo *et al.*, 2006; Mallick *et al.*, 2007; Seo *et al.*, 2007) (Figure 14). Ces études ont montré par exemple l'existence de différentes dioxygénations initiales (en positions 1,2 et en positions 3,4) et de clivages possibles en position *meta* et *ortho* des dihydroxyphénanthrènes obtenus. La souche *Alcaligenes faecalis* AFK2, quant à elle, utilise le phénanthrène comme seule source de carbone et d'énergie via le *o*-phthalate (Peng *et al.*, 2008). Chez cette souche, certains gènes non encore retrouvés chez d'autres espèces ont pu être caractérisés, comme *phnC* (codant une 3,4-dihydroxyphénanthrène dioxygénase), *phnI* (codant une 2-carboxybenzaldéhyde déshydrogénase) et *phnH* (codant une *trans*-2-carboxybenzal-pyruvate hydratase-aldolase), permettant une utilisation spécifique du phénanthrène.

Enfin, chez le genre *Sphingomonas*, il existe de nombreuses souches capables de dégrader le phénanthrène, via les mêmes étapes que celles d'autres bactéries déjà décrites (Pinyakong *et al.*, 2003a). Cependant, il faut noter que, contrairement au genre *Pseudomonas*, les gènes impliqués dans la dégradation des HAP sont dispersés au sein de différents opérons contenant également des gènes impliqués dans la dégradation de composés monoaromatiques, portés généralement par des plasmides (Figure 13) (Kim et Zylstra, 1999; Romine *et al.*, 1999b; Pinyakong *et al.*, 2003a; Demaneche *et al.*, 2004; Ní Chadhain *et al.*, 2007; Yu *et al.*, 2007). L'action de ces différentes enzymes va permettre la formation de 1-hydroxy-2-naphthoaldéhyde qui va ensuite donner l'acide 1-hydroxy-2-naphthoïque et intégrer la voie de Evans (Cho *et al.*, 2005).

7.3. Les HAP à quatre cycles : exemple du fluoranthène

En ce qui concerne le fluoranthène, plusieurs espèces appartenant notamment aux genres *Alcaligenes, Pasteurella, Sphingomonas, Mycobacterium* et *Burkholderia* sont connus pour le dégrader (Tableau 3). Les études les plus complètes ont été réalisées sur le genre *Mycobacterium*, de par ses capacités à dégrader les HAP avec un grand nombre de cycles, et donc les plus récalcitrants (comme le benzo[a]pyrène) (Figure 15) (Kim *et al.*, 2006; Kweon *et al.*, 2007; Lee *et al.*, 2007). Elles montrent que la dégradation du fluoranthène nécessite les mêmes étapes biochimiques que la dégradation du naphtalène et du phénanthrène. Cependant, beaucoup de métabolites intermédiaires ont pu être détectés, constituant parfois des impasses métaboliques (Figure 15).

La dégradation bactérienne principale du fluoranthène chez les espèces du genre *Mycobacterium* est réalisée par une dioxygénase initiale formant divers dihydrodiols (Figure 15) (Rehmann *et al.*, 2001; Lee *et al.*, 2007). Ces dihydrodiols sont ensuite convertis pour

formylacénaphtylène-1-carboxylique ; (**18**) acide acénapthylène-1,2-dicarboxylique ; (**19**) acénapthylène ; (**20**) 1,2-dihydroxyacénaphtylène ; (**21**) acide 2-(méthoxyméthyl)-acénapthylène-1-carboxylique ; (**22**) acide 2-one-acénapthylène-1-carboxylique ; (**23**) acénapthylène-1(2*H*)-one ; (**24**) 1*H*,3*H*-benzo(*de*)isochromèn-1-one ; (**25**) acide naphtalène-1,8-dicarboxylique ; (**26**) acide *cis*-1-hydroxyacénaphtylène-2-carboxylique ; (**27**) acénapthylèn-1(2*H*)-one ; (**28**) acénapthylèn-1-ol ; (**29**) acénapthoquinone ; (**30**) 9-fluorénone-1(carboxy-2hydroxy-propénol) ; (**31**) acide 9-fluorénone-1carboxylique ; (**32**) 9-fluorénol ; (**33**) 9-fluorénone ; (**34**) acide 9-carboxyméthylène-fluorène-1carboxylique ; (**35**) acide 2',3'-dihydroxybiphényle-2,3-dicarboxylique ; (**36**) 1-indanone ; (**37**) acide 2-hydroxy-4-(2-oxo-acénaphtène-1-yl)-but-3-énoïque ; (**38**) acide benzène-1,2,3-carboxylique ; (**39**) acide 2,3-dihydroxy-benzoïque ; (**40**) 2-carboxybenzaldéhyde ; **(41**) *o*-phthalate (adaptée de Seo *et al*, 2009).

former les dixydroxyfluoranthènes respectifs (Lee *et al.*, 2007). Ces derniers peuvent parfois former des diméthoxyfluoranthènes comme ceux détectés chez la souche *Mycobacterium* sp. JS14 (Lee *et al.*, 2007) qui seront également dégradés. Tous ces composés vont conduire à la formation du *o*-phthalate, dans certains cas en passant par la formation d'un intermédiaire : l'acide naphtalène-1,8-dicarboxylique (Figure 15) (Rehmann *et al.*, 2001; Kweon *et al.*, 2007; Lee *et al.*, 2007).

Chez *Mycobacterium* sp. JS14, 25 protéines ont été mises en évidence par électrophorèse bidimensionnelle, pour la plupart connues comme impliquées dans la dégradation du biphényle (Lee *et al.*, 2007). Parmi ces 25 protéines, il a été possible d'identifier deux sous-unités d'oxygénase initiales, nommées PhtA et PhtB, une ferrédoxine réductase, nommée BphG, et une biphényle dioxygénase nommée BphC). Une autre étude regroupant des données de métabolomique, de génomique et de protéomique portant sur la souche *Mycobacterium vanbaalenii* PYR-1 a également mis en évidence 53 protéines potentiellement impliquées dans la dégradation du fluoranthène. Les gènes codant ces protéines sont regroupées dans trois opérons différents (Kweon *et al.*, 2007). A partir de ces données, il a été montré que la protéine la plus exprimée est une dioxygénase initiale (NidA3 et NidB3). L'activité de cette protéine a été vérifiée par expression hétérologue chez la souche *E. coli*, où l'apparition de dihydrodiols a été mesurée pour la plupart des HAP testés, indiquant bien que c'est cette dernière qui est responsable de l'attaque initiale sur les HAP (Kim *et al.*, 2006).

Des études sur la souche *Sphingomonas paucimobilis* sp. EPA505 ont également mis en évidence une autre voie potentielle (Story *et al.*, 2000; Story *et al.*, 2001; Story *et al.*, 2004). En effet, grâce à des mutants aléatoires portant des insertions Tn5, plusieurs métabolites de dégradation du fluoranthène ont été isolés. Il s'agit du *cis*-9,10-fluoranthène dihydrodiol, du 9,10-dihydroxyfluoranthène, de l'acide 2-hydroxy-4-(2-oxo-acénaphtène-1-yl)-but-3-énoïque, de l'acide hydroxyacénaphtoïque, de l'acétonaphténone, de l'acétonaphtoquinone, de l'acide naphtalène-1,8-dicarboxylique et enfin de l'acide 2,3-dihydroxy-benzoïque, métabolite final rejoignant la voie de dégradation *meta* (Figure 15) (Story *et al.*, 2001; Pinyakong *et al.*, 2003a). L'étude de ces mutants a également mis en évidence plusieurs gènes putatifs indispensables à la dégradation du fluoranthène, comme *pbhA* (codant une 2,3-dihydroxybiphényle 1,2-dioxygénase), *pbhB* (codant une sous-unité de ferrédoxine) et *pbhC* (codant une *trans*-o-hydroxybenzylidène-pyruvate hydratase-aldolase) (Story *et al.*, 2000).

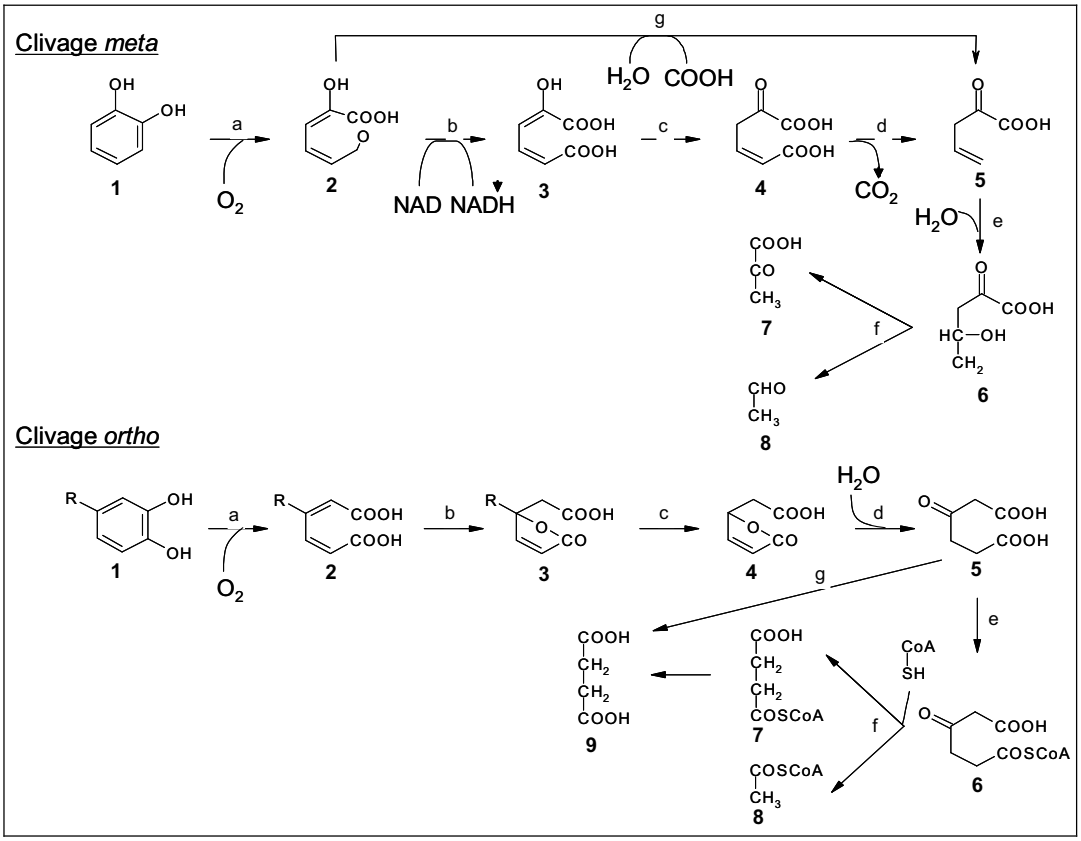


Figure 16 : Voies basses de dégradation du catéchol et du protocatéchuate, dite voies de clivage *meta* et *ortho*.

Voie de clivage *meta*.

Composés : (**1**) Catéchol ; (**2**) 2-hydroxymuconate semi aldéhyde ; (**3**) 4-oxalocronate ; (**4**) 2-oxo-hex-3-ène-1,6-dionate ; (**5**) 2-oxopent-4-ènoate ; (**6**) 4-hydroxy-2-oxovalérate ; (**7**) pyruvate ; (**8**) acétaldéhyde.

Enzymes et nomenclature des gènes correspondants donnée pour le plasmide pWWO de *Pseudomonas putida* mt-2) : (a) Catéchol 2,3-dioxygénase (*xylE*) ; (b) 2-hydroxymuconate semialdéhyde déshydrogénase (*xylG*) ; (c) 4-oxalocronate isomérase (*xylH*) ; (d) 4-oxalocronate décarboxylase (*xylI*) ; (e) 2-oxopent-4-ènoate hydratase (*xylJ*) ; (f) 4-hydroxy-2-oxovalérate aldolase (*xylK*) ; (g) 2-hydroxymuconate semialdéhyde hydrolase (*xylF*).

Voie de clivage *ortho*.

Composés : (**1**) Catéchol (R=H) ou protocatéchuate (R=COOH) ; (**2**) muconate ou 3-carboxymuconate ; (**3**) muconolactone ou 3-carboxymuconolactone ; (**4**) 3-oxoadipate énol lactone ; (**5**) 3-oxoadipate ; (**6**) 3-oxodipyl-CoA ; (**7**) succinyl-CoA ; (**8**) acétyl-CoA ; (**9**) succinate. Formation de CO_2 entre (**3**) et (**4**) si R=COOH.

Enzymes et nomenclature des gènes correspondants donnée chez *Pseudomonas putida* PRS1 et KT2240 : (a) Catéchol 1,2-dioxygénase (*catA*) ou protocatéchuate 3,4-dioxygénase (*pcaGH*) ; (b) enzyme de lactonisation du *cis*-*cis*-muconate (cycloisomérase) (*catB*) ou enzyme de lactonisation du carboxy-*cis*-*cis*-muconate *(pcaB*) ; (c) muconolactone isomérase (*catC*) ou 3-carboxy-muconolactone décarboxylase (*pcaC*) ; (d) 3-oxoadipate énol lactone hydrolase (*catD, pcaD*) ; (e) 3-oxoadipate : succinyl-CoA transférase (*catIJ, pcaIJ*) (f) 3-oxoadipyl-CoA-thiolase (*catF, pcaF*) ; (g) 3-oxoadipate : succinyl-CoA transférase (*catIJ, pcaIJ*) (adaptée de Vandecasteele, 2005).

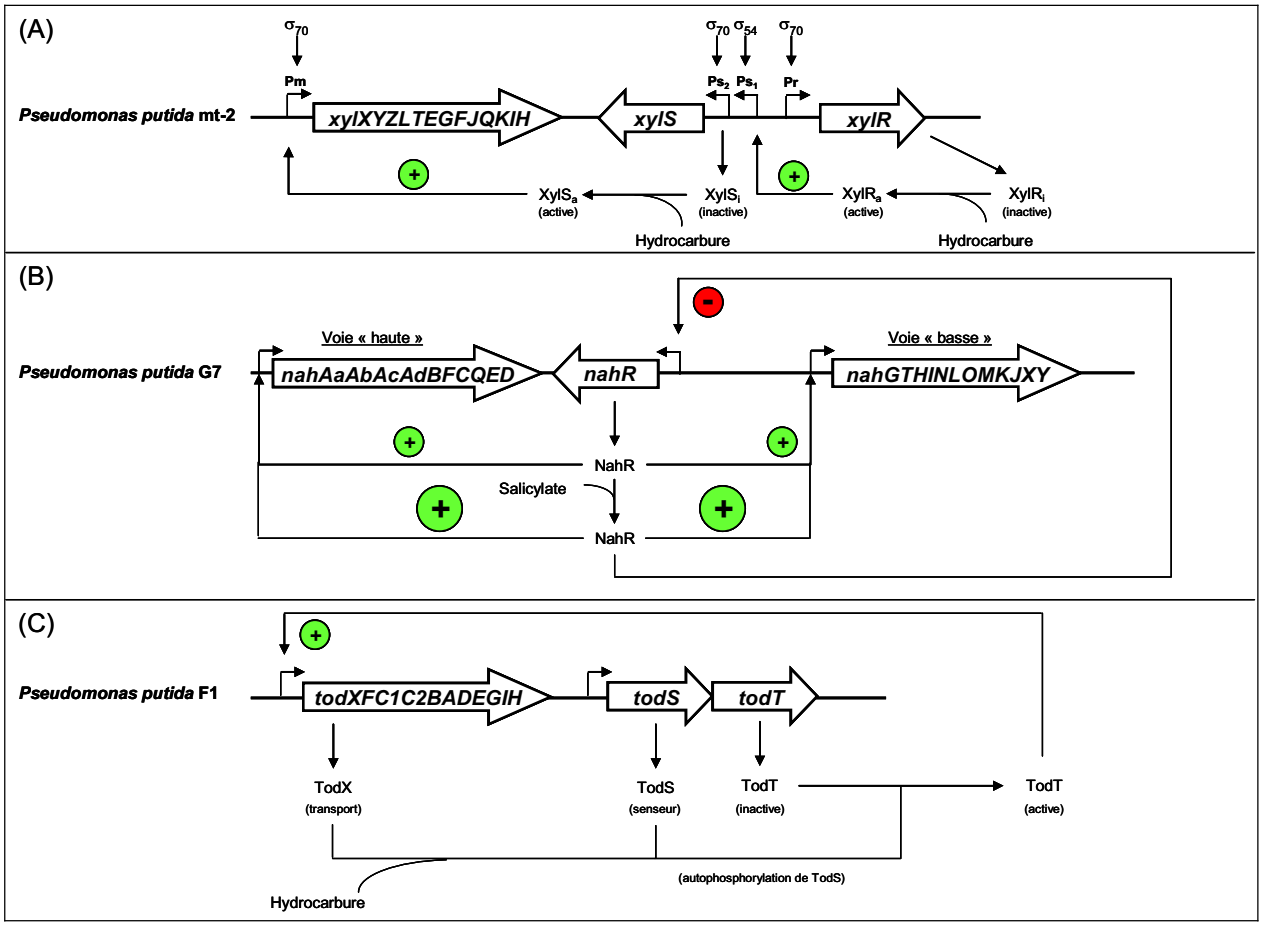
7.4. La production d'énergie via : les voies *ortho*, *meta* ou du gentisate

Comme détaillé précédemment, le catéchol, le protocatéchuate et le gentisate sont des molécules intermédiaires clés dans la dégradation bactérienne aérobie des HAP (Barnsley, 1976; Vandecasteele, 2005). Ainsi, bon nombre d'hydrocarbures aromatiques polycycliques (comme le naphtalène ou le phénanthrène), et de produits aromatiques monosubstitués et 1,2-disubstitués, (tels que le benzoate, le salicylate et l'anthranilate) sont convertis en ces métabolites. La dégradation du catéchol a été fortement décrite dans de nombreuses revues et études (Barnsley, 1976; Habe et Omori, 2003; Pinyakong *et al.*, 2003a; Peng *et al.*, 2008; Seo *et al.*, 2009). Elle peut se faire via deux phases distinctes : la voie dite *meta*, et la voie dite *ortho* (Figure 16). Il est intéressant de noter que certaines souches possèdent les deux voies basses *meta* et *ortho* de dégradation. C'est le cas de *Rhodococcus* sp. DK17 (Kim *et al.*, 2002), et de *Pseudomonas putida* F1 qui peuvent utiliser le clivage *ortho* pour le catabolisme spécifique du benzène par exemple (Habe et Omori, 2003), alors que pour la plupart des autres molécules aromatiques, le clivage *meta* est réalisé (c'est le cas des méthylbenzènes ou du *p*-xylène).

7.4.1. La voie de clivage dite meta : organisation génétique et régulation

Cette voie est initiée par le clivage du catéchol à l'aide d'une catéchol-2,3-dioxygénase (codée par *nahH* porté par le plasmide NAH7 chez *Pseudomonas putida* G7, ou *xylE* porté par le plasmide TOL chez *Pseudomonas putida* mt-2), et se termine par la formation de pyruvate et d'acétaldéhyde (Figure 16). Il est intéressant de noter que la voie de clivage comporte deux alternatives chez les deux souches citées précédemment : soit une voie directe en passant par une 2-hydroxymuconate semialdéhyde hydrolase (codée par *nahN*), soit une branche annexe comportant une déshydrogénase, une isomérase et une décarboxylase (codées respectivement par *nahI, nahJ* et *nahK*) (Habe et Omori, 2003).

Chez *P. putida* mt-2, les gènes impliqués dans la voie de clivage (13 gènes : *xylXYZLTEGFJQKIH*) sont organisés en un opéron appelé *meta* au sein d'un plasmide TOL (Spooner *et al.*, 1986; Marques *et al.*, 1998). La régulation de cet opéron est contrôlé par deux gènes adjacents : *xylS* et *xylR*, régulateurs de type AraC/XylS (Figure 17, A page suivante) (Spooner *et al.*, 1986; Marques *et al.*, 1998). Ces derniers agissent en tant qu'activateurs de transcription en présence d'effecteurs (dans notre cas, des hydrocarbures), et en absence de ces effecteurs, les protéines obtenues sont inactives, les gènes de l'opéron sont alors faiblement exprimés (Figure 17, A) (Marques *et al.*, 1998; Vandecasteele, 2005).



<u>Figure 17 :</u> Régulations de l'expression des gènes codant les protéines impliquées dans les voies de clivage dite *meta*.

(A) Régulation de l'opéron appelé *meta* chez *Pseudomonas putida* mt-2 (Spooner *et al.*, 1986; Marques *et al.*, 1998). Le gène *xylS*, sous le contrôle du promoteur Ps_2, produit faiblement une enzyme inactive $XylS_i$. Cette dernière, en présence d'hydrocarbure, devient active, et entraîne l'activation de l'expression de l'opéron *meta*. La protéine XylR, rendue également active par la présence d'effecteurs, entraîne l'expression du gène *xylS* sous le contrôle d'un second promoteur, Ps_1. (B) Régulation des opérons codant pour les protéines des voies de dégradation hautes et basses chez *Pseudomonas putida* G7 (You *et al.*, 1988). La protéine NahR, de la famille des régulateurs transcriptionnels de la famille LysR, est exprimée de manière constitutive, et active l'expression des deux opérons. En présence de salicylate, NahR interagit avec ce composé, active fortement l'expression des deux opérons, et réprime sa propre expression. (C) Régulation de l'opéron appelé *tod* chez *Pseudomonas putida* F1 (Choi *et al.*, 2003). La régulation de l'opéron *tod* s'appuie sur deux protéines : un senseur appelé TodS, et un régulateur appelé TodT. Associé à une protéine de transport appelé TodX, TodS détecte la présence d'effecteurs (comme du toluène par exemple). TodS va alors s'autophosphoryler, interagir avec TodT et rendre TodT active. Cette dernière va alors activer l'expression de l'opéron *tod*.

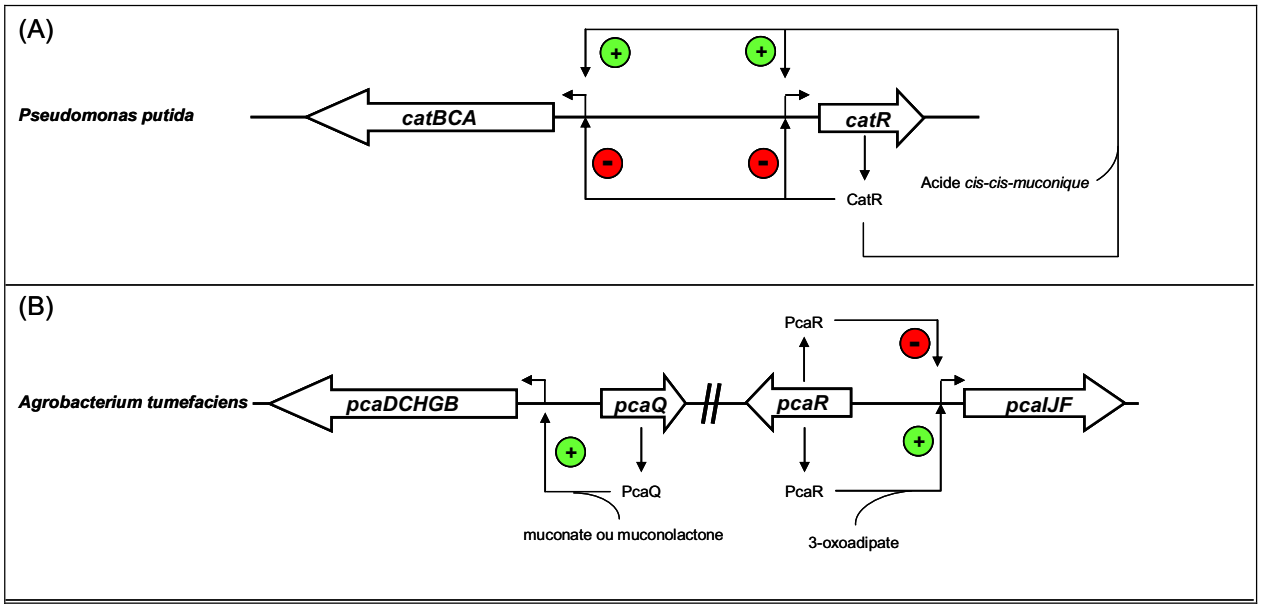
Un autre cas de régulation de la voie *meta* existe au sein du plasmide NAH7 de *Pseudomonas putida* G7, où l'on trouve 11 gènes allant de *nahG* à *nahY* proches de ceux décrits pour *P. putida* mt-2 (Figure 17, B). La même organisation génétique a été démontrée dans d'autres plasmides, comme pND6-1 chez *Pseudomonas* sp. ND6, pDTG1 chez *P. putida* NCIB9816-4, pKA1 chez *P. fluorescens* 5R ainsi qu'au sein des chromosomes de *P. stutzeri* AN10, *Pseudomonas* sp. S47, *P. putida* OUS82 et *P. aeruginosa* PaK1 (Habe et Omori, 2003). Cet opéron est contrôlé par NahR codée par le gène *nahR* (You *et al.*, 1988). Dans notre cas, la protéine régulatrice NahR est exprimée de manière constitutive et active l'expression de deux opérons, celui spécifique à la dégradation du naphtalène (« voie haute ») et celui de la voie dite *meta* (ou « voie basse ») (Figure 17, B). Néanmoins, en présence de salicylate (un intermédiaire entre ces deux voies), NahR interagit avec ce composé et active fortement l'expression des deux opérons, mais réprime également sa propre expression (Schell et Poser, 1989).

Un autre système de régulation a été décrit chez *P. putida* F1, pour l'opéron (*todXFC1C2BADEGIH*) codant les enzymes responsables des voies dites « haute » et « basse ». Cette souche effectue la dégradation de divers composés (comme le benzène ou le naphtalène) via la voie *tod* de dégradation (Figure 17, C) (Choi *et al.*, 2003). L'expression de cet opéron est régulée par les produits des gènes *todS* et *todT*, localisés au sein d'une même unité transcriptionnelle indépendante positionnée en aval du cluster (Choi *et al.*, 2003). Ces deux protéines agissent l'une comme senseur, et l'autre comme régulateur. TodS, en présence d'effecteur, va s'autophosphoryler et activer TodR, permettant l'expression de l'opéron *tod*.

7.4.2. La voie de clivage dite ortho : organisation génétique et régulation

La voie de clivage dite *ortho* (encore appelée la voie du β-kétoadipate) permet la dégradation du catéchol et du protocatéchuate par deux branches différentes (Figure 16). Cette voie de clivage est utilisée par de nombreuses espèces bactériennes (comme *Pseudomonas putida*, *Bulkhorderia cepacia*, *Agrobacterium tumefaciens* ou *Rhodococcus erythropolis*) (Seo *et al.*, 2009). Pour le catéchol, cette voie débute par la transformation du catéchol en acide *cis,cis*-muconique, via une catéchol-1,2-dioxygénase, et permet la formation de succinate et de l'acétyl-CoA). Pour le protocatéchuate, c'est la protocatéchuate-3,4-dioxygénase qui initie la dégradation, en formant du 3-carboxymuconate, pour former au final les mêmes produits que le catéchol (succinate, acétate et CO_2).

Les enzymes impliquées dans la voie de clivage dite *ortho* sont également codées par un opéron impliquant plusieurs gènes, contrôlés par une protéine régulatrice. Cette dernière



<u>Figure 18 :</u> Régulations de l'expression des gènes codant les protéines impliquées dans les voies de clivage dite *ortho*.

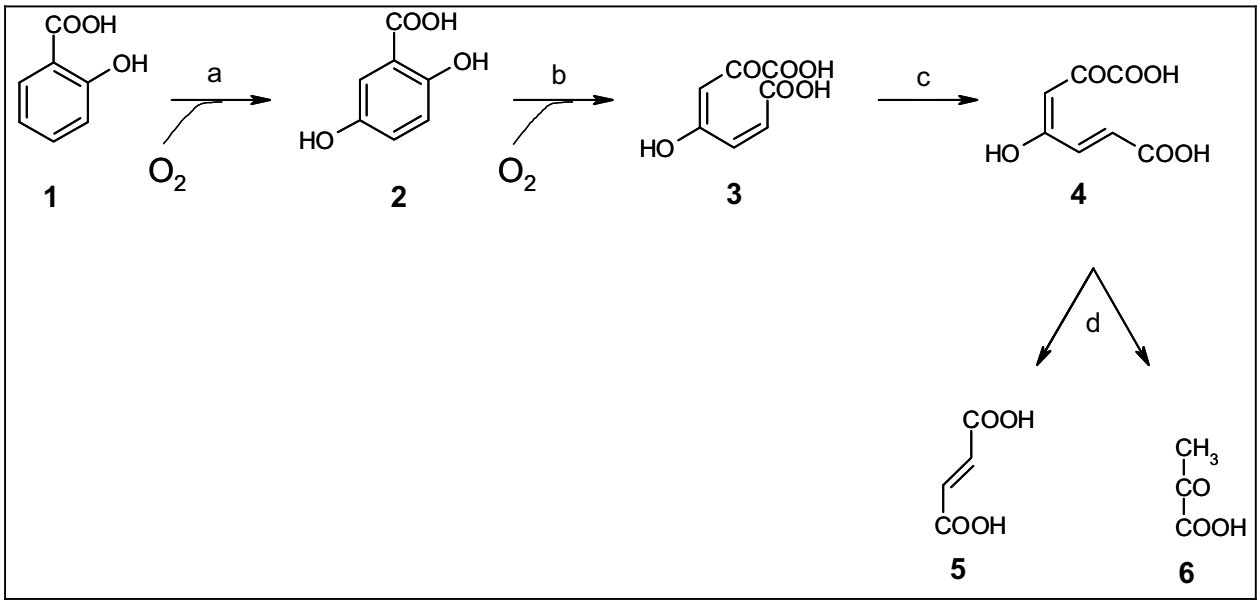
(A) Régulation de l'opéron *cat* pour la voie de dégradation *ortho* chez *Pseudomonas putida* (McFall *et al.*, 1998). La protéine CatR seule se lie aux promoteurs de l'opéron *cat* et de *catR*, empêchant la fixation de l'ADN polymérase. En présence d'acide *cis-cis*-muconique, CatR interagit avec ce composé, change de conformation et modifie sa liaison à l'ADN, levant l'inhibition. (B) Régulation des opérons impliqués dans la voie de clivage *ortho* chez *Agrobacteruim tumefaciens* (Parke, 1993; MacLean *et al.*, 2006). Chaque opéron est contrôlé par une protéine régulatrice dépendante de la présence d'un intermédiaire de dégradation. La protéine PcaQ active ainsi l'expression de *pcaDCHGB* en présence de muconate ou de muconolactone. L'expression de la seconde voie (*pcaIJF*) est, quant à elle, modulée via la protéine régulatrice PcaR (de la famille IclR). Sans effecteur, il joue le rôle de répresseur, et en présence d'effecteur comme le 3-oxoadipate, PcaR change de conformation et lève l'inhibition des gènes *pcaIJF*.

(de type LysR) est encodée par le gène *catR* pour le catéchol chez l'espèce *Pseudomonas putida* (Figure 18, A) (Rothmel *et al.*, 1990). La protéine CatR peut interagir avec les promoteurs de *catBCA* et de *catR*. En absence d'effecteur, cette protéine bloque l'expression de l'opéron. CatR, en présence d'un intermédiaire de dégradation (ici, l'acide *cis,cis*-muconique), agit en activateur de l'expression de ces gènes en modifiant sa liaison à l'ADN, et en ne bloquant plus l'ARN polymérase (McFall *et al.*, 1998).

Dans le cas du protocatéchuate, deux voies différentiellement régulées, et complémentaires sont mises en place (Figure 18, B) (Parke, 1993; MacLean *et al.*, 2006). L'activation de l'une, puis de l'autre voie va dépendre de la nature des composés formés les plus représentatifs des conditions cataboliques. Ces composés sont des métabolites de dégradation de ces voies (comme le 3-oxoadipate par exemple) (Figure 16). En effet, deux opérons existent chez certaines espèces, comme *Sinorhizobium meliloti* et *Agrobacterium tumefaciens* : *pcaDCHGB* et *pcaIJF*. Le premier est contrôlé par PcaQ, régulateur de type LysR codé par *pcaQ* (Parke, 1993; MacLean *et al.*, 2006). L'expression de la seconde voie (*pcaIJF*) est, quant à elle, modulée via la protéine régulatrice PcaR, régulateur codé par *pcaR* (Figure 18, B). Sans effecteur, il joue le rôle de répresseur, et en présence d'effecteur comme le 3-oxoadipate, PcaR change de conformation et lève l'inhibition des gènes *pcaIJF* (Parke, 1995; MacLean *et al.*, 2006). Enfin, une autre stratégie détaillée est utilisée chez *Acinetobacter* ADP1 où les deux voies de formation du catéchol et du protocatéchuate ne convergent pas. Des isoenzymes distinctes, et régulées de façon indépendante, catalysent les réactions finales de la voie de clivage *ortho* (voir Figure 16, voie de clivage *ortho*, réactions d, e, f et g) (Brzostowicz *et al.*, 2003). Cependant, on retrouve les mêmes régulateurs chez *Acinetobacter* ADP1, comme PcaU, très proche de PcaR, décrit précédemment.

7.4.3. La voie de clivage du gentisate, organisation génétique et régulation

Certains microorganismes, quant à eux, ne passent pas par le catéchol. L'acide salicylique, un métabolite terminal de la voie de dégradation dite « haute » des HAP, peut ainsi être métabolisé en gentisate (ou 2,5-dihydroxybenzoate), via une monoxygénase (composée de deux oxygénases, d'une sous-unité de ferrédoxine et d'une ferrédoxine réductase) : la salicylate-5-hydroxylase (codée par quatre gènes *nagAaAbGH* chez *Polaromonas naphthalenivorans* CJ2 ou *Ralstonia* sp. U2) qui montre de fortes similarités avec la naphtalène dioxygénase (Zhou *et al.*, 2001; Zhou *et al.*, 2002; Jeon *et al.*, 2006). Puis, le gentisate ainsi formé va être métabolisé en acide pyruvique et en fumarate, pouvant ainsi rejoindre le métabolisme central. Cette dégradation a lieu via trois étapes enzymatiques



<u>Figure 19 :</u> Voie de clivage dite du gentisate via le salicylate.

Composés : (**1**) Salicylate ; (**2**) gentisate ; (**3**) maléylpyruvate ; (**4**) fumarylpyruvate ; (**5**) pyruvate ; (**6**) fumarate.

Enzymes et nomenclature des gènes correspondants donnée chez *Polaromonas naphthalenivorans* CJ2 et *Ralstonia* sp. U2 : (a) Salicylate-5-hydroxylase (*nagAaAbGH*) ; (b) gentisate-1,2-dioxygénase (*nagI*) ; (c) maléylpyruvate isomérase (glutathione-dépendante) (*nagL*) ; (d) fumarylacétoacétate hydrolase (*nagK*) (adaptée de Vandecasteele, 2005).

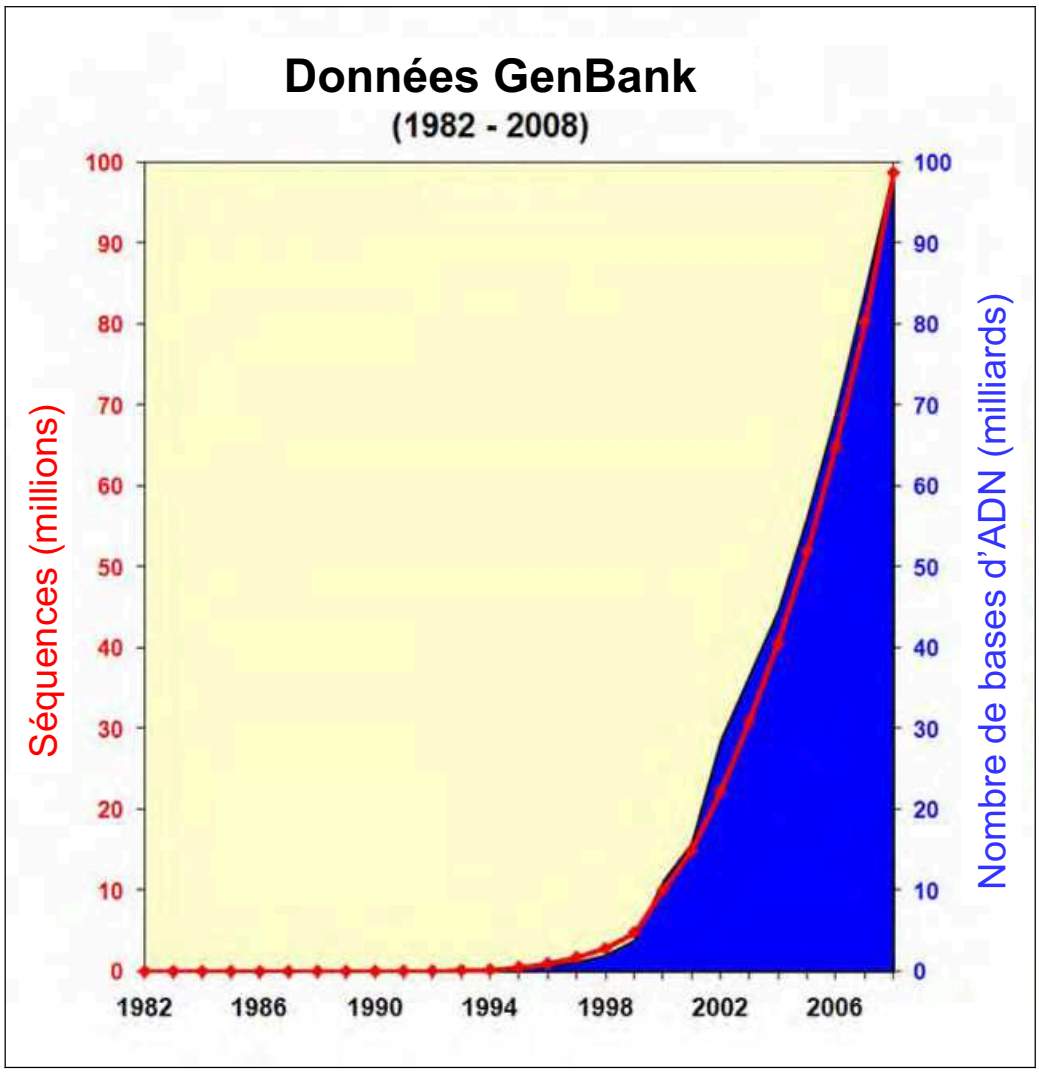
spécifiques impliquant : une dioxygénase (la gentisate-1,2-dioxygénase) réalisant la fission du cycle entre le groupement carboxylate et l'hydroxyle voisin, une isomérase et une hydrolase (Figure 19) (Zhou *et al.*, 2001).

Chez *Ralstonia* sp. U2, les gènes impliqués dans la dégradation du gentisate sont *nagIKL*, codant respectivement pour une dioxygénase, une isomérase et une hydrolase (Zhou *et al.*, 2001). L'opéron *nagIKL* est sous la régulation de la protéine NagR, qui, en présence de salicylate, induit son expression (Jones *et al.*, 2003). Chez *Polaromonas naphthalenivorans* CJ2, les gènes *nag* sont regroupés au sein de deux opérons (Jeon *et al.*, 2006). Le plus petit opéron (*nagR*$_2$ORF2*I*$_2$*KL*) semble régulé par la protéine NagR$_2$ de type MarR (Jones *et al.*, 2003) qui n'est ni induite par le salicylate, ni par le gentisate, ni par le naphtalène. De plus, cette souche possède deux copies de la gentisate-1,2-dioxygénase. L'effecteur du cluster portant la deuxième copie semble ainsi encore inconnu.

8. Conclusion

Les informations disponibles sur les enzymes impliquées dans les voies de dégradation des HAP sont assez fragmentaires, sauf pour les HAP modèles comme le phénanthrène ou le naphtalène (Seo *et al.*, 2009). De plus, celles-ci ne sont disponibles que pour des microorganismes généralement cultivables et isolés. Or les microorganismes épurateurs agissent le plus souvent en *consortia* au sein des sols, et beaucoup d'entre eux ne peuvent être isolés et étudiés facilement. De plus, ces *consortia* possèdent des capacités de dégradation plus importantes que les microorganismes isolés (Molina *et al.*, 2009; Vila *et al.*, 2010). Obtenir plus d'informations sur ces *consortia* permettrait de franchir une nouvelle étape dans la compréhension des capacités de biodégradation des HAP, et donc d'améliorer les techniques de bioremédiation.

C'est pourquoi il est devenu indispensable d'étudier les écosystèmes dans leur globalité. Ainsi, les projets de séquençage systématique de métagénomes ont permis d'acquérir de nombreuses informations de séquences issues d'environnements complexes. Cependant, la grande quantité de données ainsi produite nécessite maintenant de développer des outils informatiques permettant de réaliser une annotation fonctionnelle correcte, et une fouille de données efficace et rapide. Le domaine de la bioinformatique s'est donc concentré depuis quelques années au développement d'outils informatiques répondant à ces besoins. Ils permettent ainsi la consultation et/ou la reconstruction de processus biologiques, afin de mieux appréhender les potentialités biologiques de ces écosystèmes dans leur globalité.



<u>Figure 20 :</u> Evolution des données soumises au sein de la base de séquences GenBank de 1982 à 2008.
Source (GenBank) : http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html

Chapitre III : Fouille des données de génomique pour la consultation et la reconstruction de voies métaboliques

1. L'explosion des données de génomique

Le séquençage conventionnel d'ADN repose sur la technique développée par Sanger en 1977 (Sanger *et al.*, 1977). Celle-ci a bénéficié par la suite de nombreuses améliorations (fluorophores, migration en capillaire) permettant notamment son automatisation. Cette technique a été utilisée pour assurer le séquençage de génomes complets, le premier ayant été celui d'*Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995). Plus récemment, de nouvelles techniques de séquençage dites de très haut-débit, comme le pyroséquençage ont été développées (Ronaghi, 2001). Elles ont facilité la mise en place de plateformes assurant le séquençage rapide de génomes, voire de métagénomes complets (Margulies *et al.*, 2005).

Ces nouvelles avancées techniques ont permis la multiplication des projets de séquençage systématique, entraînant une explosion du nombre de séquences dans les bases de données (Figure 20). La banque de données GenBank est ainsi passée de 39 533 séquences (pour un total de 49 179 285 bases) en 1990, à 119 112 251 séquences (pour un total de 114 348 888 771 bases) en 2010. La masse de données générée empêche cependant une vérification expérimentale systématique du produit codé par les gènes localisés au sein de ces séquences. Les approches *in silico* sont ainsi préférentiellement utilisées pour réaliser l'annotation fonctionnelle des séquences stockées au sein de ces banques de données internationales. La complexité des bases de données peut rendre difficile la recherche d'informations pertinentes. De même, la reconstruction métabolique *in silico* d'environnements demeure très complexe même si de nouvelles approches ont été validées sur la base de génomes complètement séquencés.

2. Annotation fonctionnelle *in silico* des données de génomique

L'annotation fonctionnelle *in silico* des données de génomique repose sur la relation de similarité entre séquences. Il est ainsi possible de retrouver pour une séquence inconnue des séquences proches dont la fonction a été définie. Pour améliorer la qualité de ces annotations fonctionnelles, des approches de phylogénie moléculaire ont été développées, assurant

notamment une meilleure identification des séquences orthologues. Pour assurer ce type d'annotation, il est aussi envisageable d'exploiter les données d'organisation des génomes (synténie, prédiction d'opérons).

2.1. Annotation fonctionnelle par comparaison de séquences

L'annotation fonctionnelle permet d'attribuer une fonction aux produits des gènes, et est basée essentiellement sur la recherche de similarité entre séquences nucléiques et/ou protéiques, qu'elles soient complètes ou non.

2.1.1. Annotation fonctionnelle par similarité de séquences primaires

Pour comparer deux séquences, il est nécessaire de procéder à leur alignement, afin de repérer les zones similaires qu'elles peuvent partager. Deux approches peuvent être exploitées, selon le degré de conservation des gènes étudiés. L'alignement global (Needleman et Wunsch, 1970), va permettre de comparer deux séquences sur toute leur longueur. L'alignement local (Waterman, 1984), quant à lui, s'affranchit des parties les plus variables des séquences, en ne prenant en compte que les régions les plus conservées. Ces deux approches s'appuient sur des matrices, développées pour calculer un score reflétant la similarité, ou les divergences entre séquences. Les matrices les plus utilisées pour calculer ces scores sont de type BLOSUM ou PAM (Wilbur, 1985; Henikoff et Henikoff, 1993). Par exemple, l'outil FASTA (Pearson et Lipman, 1988) implémente une de ces deux méthodes (Waterman, 1984). Bien que ces deux précédents outils soient les plus utilisés par la communauté scientifique, il existe également d'autres logiciels pour réaliser ce type de recherche comme needle (alignement global) et water (alignement local) de la suite EMBOSS (Rice *et al.*, 2000).

Ces approches vont donc permettre de rechercher au sein des bases de données, la ou les séquences similaires à la séquence requête. Cependant les masses de données contre lesquelles sont faites les comparaisons étant de plus en plus importantes, des approches probabilistes (ou heuristiques) sont employées pour limiter les temps de traitement. C'est le cas des outils dérivant de l'algorithme BLAST. (Altschul *et al.*, 1990). Sa rapidité d'exécution est notamment l'une des raisons pour lesquelles les outils BLAST sont devenus des références pour la recherche de similarité entre séquences.

Dans la plupart des cas, les recherches de similarités sont effectuées contre les bases de données internationales. On peut ainsi citer les bases de séquences nucléiques comme EMBL (European Molecular Biology Laboratory) (http://www.ebi.ac.uk/embl/), GenBank (http://www.ncbi.nlm.nih.gov/) ou encore DDBJ (DNA Data Bank of Japan)

(http://www.ddbj.nig.ac.jp/) (Cochrane et Galperin, 2010). Il est également possible d'effectuer des recherches sur des banques de données protéiques comme Swiss-Prot, TrEMBL (regroupées dans UNIPROT : http://www.expasy.ch/sprot/), ou encore GenPept (http://pbil.univ-lyon1.fr/pf_bioinfo/article202.html) (Cochrane et Galperin, 2010). La base de données Swiss-Prot est la banque présentant la plus haute valeur ajoutée. En effet, cette base contient des séquences non redondantes, dont l'annotation a été validée par des experts. La base TrEMBL contient, quant à elle, des données potentiellement redondantes, et la plupart du temps non vérifiées expérimentalement, puisque les séquences qu'elle regroupe sont générées automatiquement par traduction des régions codantes des séquences nucléiques de la banque EMBL.

Ces approches basées sur la similarité entre séquences permettent de définir *in silico* les fonctions des produits des gènes. Néanmoins, l'absence quasi systématique de vérification expérimentale peut conduire à l'affectation d'une fonction erronée pour le produit d'un gène. Ce type d'erreur porte sur la caractérisation d'un domaine protéique conservé entre deux séquences (similarité locale), sans tenir compte des séquences dans leur globalité. Ainsi, ces deux protéines peuvent avoir des fonctions biologiques différentes, tout en partageant un même domaine protéique conservé (Gerlt et Babbitt, 2000).

2.1.2. Annotation fonctionnelle par recherche de motifs ou de domaines

L'analyse des séquences et des structures protéiques a permis de constater qu'elles s'organisaient en domaines, c'est-à-dire en parties acquérant une structure, et remplissant une fonction, indépendamment du reste de la protéine. Ces domaines sont de plus, fortement conservés, de par leur rôle crucial dans la fonction biologique des protéines. Ces dernières sont en fait des « mosaïques » de ces domaines. Ainsi, avec des associations différentes de domaines, ou de motifs, il est possible de générer toute la diversité des protéines. D'où l'intérêt d'utiliser ces régions conservées pour faire de l'annotation fonctionnelle de protéines. Cependant, il faut au préalable identifier ces régions conservées, en réalisant des alignements multiples de séquences protéiques.

Deux approches ont été développées pour la génération d'un alignement multiple, avec pour point de départ un alignement local ou global entre deux séquences (Needleman et Wunsch, 1970; Waterman, 1984). Une première approche consiste à enrichir progressivement l'alignement obtenu, par l'ajout d'une nouvelle séquence. La seconde approche, quant à elle, va reposer sur l'identification de sous-familles de séquences. Elles seront alignées au sein d'une même sous-famille, puis les différentes sous-familles alignées entre elles. De nombreux

Tableau 4 : Banques de motifs fonctionnels les plus généralement utilisées.
(Source : annotation fonctionnelle des génomes, de la séquence à la biologie. Céline BROCHIER, Guy PERRIERE - http://biologie.univ-mrs.fr/upload/p213/annotation2007.pdf).

Nom de la banque	Source des données	Construction	Référence
PROSITE	Swiss-Prot	Expressions régulières (patterns)	(Sigrist *et al.*, 2010)
BLOCKS	UniProt	Motifs pondérés	(Pietrokovski *et al.*, 1996)
PRINTS	UniProt	Motifs alignés	(Attwood, 2002)
eMOTIFS	BLOCKS/PRINTS	Expressions régulières floues	(Cochrane et Galperin, 2010)
Profile	UniProt	Matrices de pondération (profils)	(Cochrane et Galperin, 2010)

Tableau 5 : Banques de domaines fonctionnels les plus généralement utilisées.
(Source : annotation fonctionnelle des génomes, de la séquence à la biologie. Céline BROCHIER, Guy PERRIERE - http://biologie.univ-mrs.fr/upload/p213/annotation2007.pdf).

Nom de la banque	Description des données	Référence
CDD	Alignements multiples représentant les domaines protéiques conservés de plusieurs bases (SMART, Pfam, COGs, PRK et TIGRFAMs).	(Marchler-Bauer *et al.*, 2009)
EVEREST	Domaines et familles de domaines (chaînes de Markov cachées) basés sur UniProt et PDB.	(Portugaly *et al.*, 2007)
InterDom	Domaines d'interactions protéiques putatifs.	(Ng *et al.*, 2003)
PANDIT	Alignements multiples représentant les domaines protéiques basés sur PFAM-A (familles vérifiées).	(Whelan *et al.*, 2003)
Pfam	Alignements multiples et profils pour ces derniers.	(Finn *et al.*, 2010)
ProDom	Domaines basés sur UniProt, recherche réalisée par PSI-BLAST.	(Bru *et al.*, 2005)
SBASE	Domaines protéiques.	(Vlahovicek *et al.*, 2005)
SMART	Domaines protéiques vérifiés manuellement.	(Letunic *et al.*, 2009)
TIGRFAMs	Alignements multiples représentant les domaines protéiques conservés et profiles.	(Haft *et al.*, 2003)

outils implémentent ces approches comme MULTALIN (Corpet, 1988), DIALIGN (Morgenstern *et al.*, 1998), Muscle (Edgar, 2004) et CLUSTAL (Thompson *et al.*, 1994). Le logiciel CLUSTAL (pour CLUSter ALignment) est basé sur l'utilisation d'un algorithme d'alignement progressif. Dans un premier temps, les alignements de toutes les paires de séquences sont réalisés puis, une matrice de distances de similitude est calculée afin de pondérer ces alignements individuels. Les séquences les plus proches sont ensuite alignées, puis l'alignement multiple obtenu s'enrichit progressivement avec les séquences de plus en plus distantes (Thompson *et al.*, 1994).

Les alignements multiples réalisés à partir de protéines présentant la même fonction biologique servent ensuite de support pour des méthodes mathématiques permettant la caractérisation des motifs ou domaines, au sein des séquences étudiées. Les motifs sont généralement des séquences très courtes, avec des résidus essentiels (pas nécessairement consécutifs) à la fonction biologique (catalyse, fixation d'un ligand, régulation, etc.…). Les domaines, quant à eux, sont des fragments de séquences (ou blocs) conservés entre différentes protéines. Dans certains cas, un même motif ou domaine peut être présent au sein de deux protéines n'ayant pas la même fonction biologique. Ces informations (motifs ou domaines) sont définies selon différentes approches (Cochrane et Galperin, 2010) : (*i*) par des « expressions régulières » ou des « patterns » (résidus essentiels retrouvés pour le motif et parfois séparés par des séquences peptidiques moins conservées) définis en se basant sur des alignements multiples de séquences très proches; (*ii*) par la détermination de matrices de poids (appelées des profils), dérivées des alignements multiples de séquences ; (*iii*) par la construction de modèles probabilistes pour générer des profils (chaînes de Markov cachées). L'utilisation de ces approches a ainsi permis d'identifier de nombreux motifs ou domaines fonctionnels, stockés et accessibles au sein de différentes bases de données (Tableaux 4 et 5).

La base de données de motifs PROSITE, par exemple, peut être considérée comme un dictionnaire de motifs protéiques ayant une signification biologique (Sigrist *et al.*, 2010). Sa conception repose sur trois objectifs : (1) collecter le plus grand nombre de motifs, (2) regrouper les motifs hautement spécifiques, (3) fournir une documentation complète sur chaque motif. Les banques de domaines protéiques, les plus couramment utilisés sont Pfam, et ProDom (Tableau 5) (Cochrane et Galperin, 2010). La base Pfam regroupe les alignements protéiques ainsi que les domaines représentée par des chaînes de Markov cachées, stockées au sein de deux sous-banques différentes. Pfam-A regroupe tous les domaines protéiques vérifiés manuellement, et Pfam-B des domaines protéiques obtenus automatiquement. La banque ProDom regroupe des domaines générés par une approche itérative de recherche de similarité

effectuée sur des séquences de protéines. Les données de séquences provenant d'UNIPROT (Bru *et al.*, 2005) sont traitées par l'outil PSI-BLAST (ou Position-Specific Iterated BLAST) (Altschul *et al.*, 1997) qui produit alors des profils. L'approche implémentée dans PSI-BLAST se base sur la construction progressive d'une matrice qui traduit la probabilité d'observer un ou plusieurs acides aminés pour chaque site moléculaire. Cette matrice est améliorée à chaque étape de recherche avec des séquences supplémentaires. Ainsi, il est possible de détecter des séquences beaucoup plus divergentes qu'avec une approche reposant sur le simple algorithme BLAST.

Les nombreuses banques de motifs ou de domaines protéiques sont désormais interrogeables par le *meta* système de recherche InterProScan (Hunter *et al.*, 2009). Il permet d'avoir accès à la base de données composite InterPro, qui englobe les données de différentes banques de signatures protéiques (Gene3D, HAMAP, PANTHER, PIRSF, PRINTS, PROSITE, Pfam, ProDom, SMART, SUPERFAMILY et TIGRFAMs) (Hunter *et al.*, 2009). Parallèlement à cette recherche, InterProScan permet aussi de décrire la ou les fonction(s) de la protéine requête, selon des termes contrôlés du « Gene Ontology » (ou GO).

2.1.3. Description de la fonction biologique : Gene Ontology

Le « Gene Ontology » (GO) a été développé afin de proposer un vocabulaire standardisé limitant ainsi les erreurs d'annotation. La description proposée par GO présente trois niveaux de précision permettant de définir : la fonction moléculaire, la fonction biologique et la localisation cellulaire où cette fonction est exercée (Ashburner et Lewis, 2002). Ce système permet de décrire l'ensemble des fonctions biologiques connues, tout en étant suffisamment précis pour distinguer les spécificités d'une protéine d'intérêt. Il existe un grand nombre d'outils disponibles permettant d'utiliser les données GO pour effectuer des annotations fonctionnelles. Le seul ayant été développé par le Consortium GO est AmiGO (Carbon *et al.*, 2009). Par exemple, il est possible de lui fournir une ou plusieurs séquences (nucléiques ou protéiques) pour effectuer une recherche de similarité via l'algorithme BLAST, contre la banque de séquences protéiques spécifique à GO. Les résultats obtenus vont permettre de fournir une liste de termes GO les plus adaptés par rapport à cette recherche de similarité. Il est également possible d'effectuer une annotation GO par similarité de séquence. Des logiciels similaires ont été développés comme BLAST2GO (Conesa *et al.*, 2005), ou InterProScan (Hunter *et al.*, 2009).

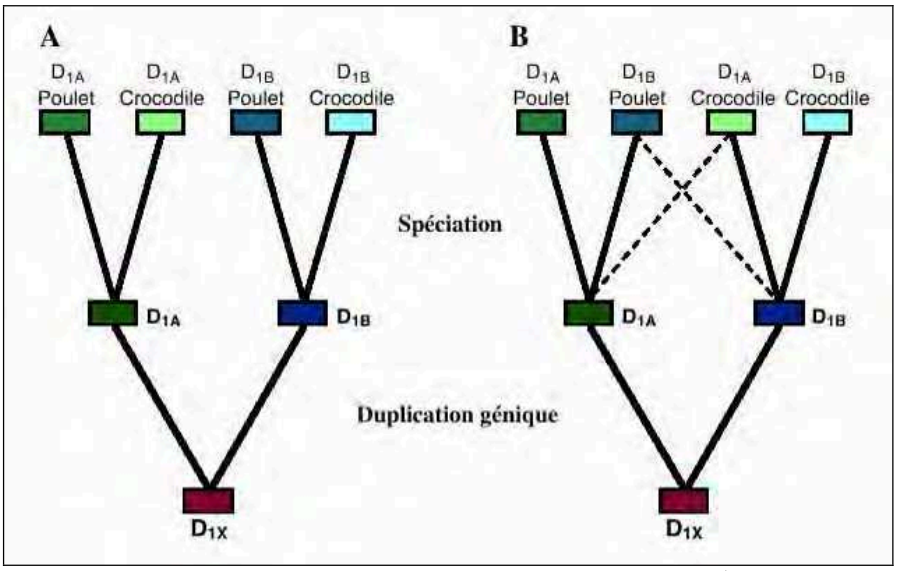


Figure 21 : Notions de gènes paralogues et orthologues. Tirée de (Vernier *et al.*, 1993).
A. : La duplication génique du gène D_{1X} donne deux paralogues D_{1A} et D_{1B}. Après obtention de ces caractères, un phénomène de spéciation intervient au cours du temps, qui maintient dans les deux espèces, poulet et crocodile, les caractères D_{1A} et D_{1B}. Les gènes D_{1A} et D_{1B} de poulet et de crocodile sont donc respectivement des orthologues.
B. : L'évolution des séquences géniques, notamment par les phénomènes de convergence, peut rendre difficile l'identification des relations phylogénétiques réelles. Ainsi, au cours de l'évolution les différentes mutations de séquences laissent supposer que les gènes D_{1A} et D_{1B} de poulet et de crocodile ne sont pas respectivement des orthologues mais des paralogues. Les relations phylogénétiques correctes sont indiquées en pointillés.

2.2. Méthodes phylogénétiques pour l'annotation fonctionnelle

La phylogénie moléculaire permet l'étude de l'évolution des organismes vivants en vue d'établir leur parenté (Brilli *et al.*, 2008; Kuzniar *et al.*, 2008; Perrière et Brochier-Armanet, 2009). Ce lien peut être établi notamment, en comparant les séquences de biomarqueurs. Cependant, cette recherche de similarité entre séquences ne permet pas d'établir précisément si elles sont homologues (deux séquences sont dites homologues si elles possèdent un ancêtre commun). En effet, deux séquences peuvent être homologues sans similarité apparente, et inversement. Des analyses complémentaires sont donc nécessaires pour définir plus précisément les liens de parenté entre les séquences.

2.2.1. Identification de séquences orthologues

Les notions d'orthologie et de paralogie sont indispensables pour étudier des séquences homologues. Des séquences orthologues sont en fait des séquences présentes dans des organismes différents, ayant évolué à partir d'une même séquence ancestrale (Figure 21). La fonction biologique portée par ces séquences est très souvent conservée au cours de l'évolution des orthologues (Lemoine *et al.*, 2007). Les séquences dites paralogues sont, quant à elles, issues d'événements de duplication au sein d'un même génome. Au cours de l'évolution, et donc des modifications indépendantes de chacune des séquences, ces paralogues peuvent évoluer différemment, et parfois vers des fonctions différentes (Kuzniar *et al.*, 2008). Un exemple de paralogues chez de nombreux microorganismes est le gène codant pour l'ARNr 16S.

Une des approches les plus utilisées pour distinguer les orthologues des paralogues est celle dite du « bi-directional best-hit », ou BBH (ou encore Best Reciprocal Hit ou BRH). Cette approche consiste en la comparaison réciproque de séquences entre organismes (Kuzniar, van Ham et al. 2008). Ainsi, la protéine *a* codée au sein du génome A (notée *a*(A)) sera un orthologue de la protéine *b* codée au sein du génome B (notée *b*(B)), si le meilleur résultat de la comparaison de *a*(A) sur B est *b*, et réciproquement. Cependant, ce type de méthode présente des limites liées aux différences de vitesse d'évolution entre les séquences étudiées, et à l'existence des transferts horizontaux de gènes (Lemoine *et al.*, 2007; Descorps-Declère *et al.*, 2008).

Cette approche a néanmoins été mise en oeuvre pour définir des clusters de séquences homologues provenant de différents organismes (Tatusov *et al.*, 1997; Brilli *et al.*, 2008; Kuzniar *et al.*, 2008). Ces données sont notamment répertoriées au sein de la banque de donnée COGs ('Clusters of Orthologous Groups'), où l'on retrouve de nombreux clusters de

Tableau 6 : Banques de séquences et/ou de clusters homologues les plus généralement utilisées.
(Source : annotation fonctionnelle des génomes, de la séquence à la biologie. Céline BROCHIER, Guy PERRIERE - http://biologie.univ-mrs.fr/upload/p213/annotation2007.pdf).

Nom de la banque	Source des données	Contenu	Référence
ProtoMap	Swiss-Prot	Tous organismes disponibles	(Yona *et al.*, 2000)
ProClass	PIR/PROSITE	Tous organismes disponibles	(Huang *et al.*, 2000)
CluSTr	UniProt	Tous organismes disponibles	(Kriventseva *et al.*, 2001)
MetaFam	Divers (10 banques)	Tous organismes disponibles	(Silverstein *et al.*, 2001)
SYSTERS	Swiss-Prot/PIR	Tous organismes disponibles	(Krause *et al.*, 2000)
COGs	Génomes	66 Génomes complets	(Makarova *et al.*, 2007)
HOBACGEN	UniProt	Bactéries et archées	(Perrière *et al.*, 2000)
HOGENOME	UniProt	Génomes complets	(Dufayard *et al.*, 2005)
RTKdb	UniProt	Récepteurs à tyrosine kinase	(Grassot *et al.*, 2003)
ABCdb	UniProt	ABC transporteurs	(Capponi *et al.*, 2001)

protéines homologues d'origines eucaryote ou procaryote (Makarova *et al.*, 2007). D'autres bases de ce type, plus ou moins spécialisées, ont également été développées (Tableau 6). Cependant, les deux séquences les plus similaires au sein d'un jeu de données ne correspondent pas forcément aux plus proches orthologues, mais uniquement aux plus proches homologues (Descorps-Declère *et al.*, 2008). Il est donc indispensable d'appliquer d'autres approches prenant mieux en compte l'évolution des séquences pour différencier au mieux les orthologues des paralogues.

2.2.2. Comparaison de séquences par l'utilisation de méthodes phylogénétiques

Afin d'évaluer les degrés de parenté entre les séquences, deux types d'approches phylogénétiques ont été développées : (i) celles basées sur les mesures de distances entre séquences prises deux à deux, qui caractérisent le nombre de substitutions de nucléotides ou d'acides aminés entre ces deux séquences ; (ii) et celles basées sur les caractères qui s'intéressent au nombre de mutations (que ce soit des substitutions, des insertions ou des délétions) qui affectent chacun des sites moléculaires de la séquence (Brilli *et al.*, 2008; Kuzniar *et al.*, 2008; Perrière et Brochier-Armanet, 2009).

2.2.2.1. Les méthodes basées sur les distances

Les approches basées sur les distances sont généralement rapides et donnent de bons résultats pour des séquences ayant une similarité élevée. Les deux méthodes principales sont : UPGMA (Unweighted Pair Group Method with Arithmetic mean), et NJ (Neighbor Joining) (Fitch et Margoliash, 1967; Saitou et Nei, 1987). UPGMA utilise une approche de regroupement séquentiel, dans laquelle les relations entre les différentes séquences sont évaluées par la détermination du taux de similarité qu'elles présentent (Fitch et Margoliash, 1967). Ensuite, les distances calculées entre chaque paire de séquences permettront la reconstruction de l'arbre phylogénétique. L'inconvénient majeur de cette méthode est qu'elle ne prend pas en compte les différentes vitesses d'évolution. En effet, chaque séquence accumule les mutations à un rythme qui lui est propre, et qui est dicté par l'intensité de la pression de sélection à laquelle elle est soumise. La méthode du Neighbor Joining, développée par Saitou et Nei tente de corriger la méthode UPGMA afin de prendre en compte ce paramètre (Saitou et Nei, 1987). En effet, la matrice de distances est construite en pondérant la variabilité de chaque séquence par son degré de similarité avec l'ensemble des séquences analysées. Ainsi, cette méthode prend en compte le facteur temps pour la divergence moyenne des séquences. Ces deux méthodes de distance sont implémentées et utilisables via divers packages informatiques comme PHYLIP (DNADIST, PROTDIST et NEIGHBOR pour

l'approche NJ) (Fink, 1986) ou encore le logiciel MEGA (implémentant les deux méthodes) (Tamura *et al.*, 2007).

2.2.2.2. Les méthodes basées sur les caractères

Ces méthodes bien que nécessitant des temps de calcul beaucoup plus importants, permettent d'augmenter la fiabilité des résultats sur des séquences présentant moins de similarité. Deux approches existent, celles dites de parcimonie, et celles dites du maximum de vraisemblance (Felsenstein, 1978, 1981). La méthode dite de parcimonie consiste à reconstruire l'arbre phylogénétique nécessitant le plus petit nombre de changements évolutifs pour aboutir aux séquences disponibles, en s'appuyant sur les hypothèses suivantes : (i) les sites évoluent indépendamment les uns des autres, et, (ii) la vitesse d'évolution est lente et constante au cours du temps (Felsenstein, 1978). Son principal défaut est de ne considérer que les sites informatifs, et donc ceux présentant une différence de séquence (insertion, délétion, mutation, etc.…), en ignorant donc les sites n'en présentant aucune. La méthode du maximum de vraisemblance, prend en compte, quant à elle, l'ensemble des divergences et similitudes d'un jeu de séquences (Felsenstein, 1981). Pour cela, elle recherche l'arbre dont la vraisemblance est maximale en s'appuyant sur un modèle d'évolution (Felsenstein, 1981). Cette méthode évalue ainsi l'ordre (parenté), et la longueur (vitesse d'évolution) des branches de l'arbre qui sont les plus probables. On retrouve notamment ces deux approches implémentées au sein d'outils disponibles dans le package PHYLIP (avec PROML et DNAML pour l'approche du maximum de vraisemblance, et DNAPARS et PROTPARS pour la méthode dite de parcimonie) (Fink, 1986).

2.2.2.3. Exemple d'application de l'utilisation de méthodes phylogénétiques

Ce type de méthode peut notamment être utile lorsque, parmi une famille protéique très large, l'annotation fonctionnelle n'a été vérifiée expérimentalement que pour une petite fraction d'entre elles. C'est le cas de la Famille 4 des glycosyl hydrolases (Hall *et al.*, 2009), où l'activité et le spectre de substrats de 24 de ces enzymes ont été vérifiés, parmi les 383 enzymes composant cette Famille 4 (Hall *et al.*, 2009). 201 séquences de ces enzymes provenant de 102 espèces de bactéries et d'archées ont été traitées par la méthode du maximum de vraisemblance, les résultats obtenus ont permis la reconstruction d'un arbre phylogénétique pour la Famille 4 des glycosyl hydrolases. Cet arbre phylogénétique a ensuite servi de base pour l'annotation fonctionnelle de 50 de ces protéines qui étaient classées de manière ambiguë (définition de l'activité trop large, protéines hypothétiques, etc.…). De plus,

au vu des résultats, neuf séquences dont les fonctions biologiques étaient incorrectes, ont été ré-annotées (Hall *et al.*, 2009).

2.2.3. Synténie et prédiction d'opérons

2.2.3.1. La synténie

La synténie correspond à la conservation de l'ordre des gènes entre différents génomes (Mahadevan et Seto, 2010). Elle se caractérise par deux types de relations : une relation de co-localisation (gènes présents sur le même chromosome), et une relation de correspondance (similarité de séquence). Chez les procaryotes, on parle de « groupe de synténie » (ou syntons) pour décrire une organisation génomique identique entre deux génomes pour des groupes de gènes similaires (Mahadevan et Seto, 2010). Les avantages de la synténie sont nombreux : (i) par exemple, elle permet l'annotation fonctionnelle des produits des gènes en se basant aussi sur des informations de structure génomique ; (ii) elle permet aussi plus facilement d'identifier les orthologues au sein d'un mélange de séquences homologues (et donc d'orienter leur classification au sein de familles dont la fonction est déjà décrite) (Mahadevan et Seto, 2010). Ces deux avantages permettent d'améliorer, ou de confirmer l'annotation fonctionnelle de ces groupes de synténie (Descorps-Declère *et al.*, 2008).

La recherche de syntons a permis le développement de banques informatiques, comme SynteBase (Lemoine *et al.*, 2007; Lemoine *et al.*, 2008) qui regroupe des données de synténie de près de 600 génomes de procaryotes. Ces données sont accessibles via une interface graphique appelée SynteView (Lemoine *et al.*, 2008), permettant la visualisation et la comparaison de génomes choisis. Ces deux éléments : SynteBase et SynteView ont été développés dans le but de permettre une annotation plus efficace et plus rapide des futurs génomes séquencés, en regroupant des données déjà traitées et vérifiées.

L'utilisation de la synténie facilite également l'annotation fonctionnelle de gènes dont la structure est conservée au sein de nombreux organismes, comme par exemple la production d'énergie, une fonction biologique indispensable (Barbe *et al.*, 2004; Lemoine *et al.*, 2008). Ainsi, la synténie a été utilisée pour l'annotation fonctionnelle du génome séquencé de la souche *Acinetobacter baylyi* ADP1, modèle intéressant pour l'étude de la compétence pour la transformation naturelle (Barbe *et al.*, 2004). La première étape, la recherche de similarité dans les bases de données, a permis d'assigner une fonction à plus de 62 % des 3 325 gènes codant des protéines qui ont été identifiés dans la séquence génomique (35 % de façon définitive, et 27 % de façon putative) de cette souche. Mais, plus de 60 % des gènes d'*Acinetobacter baylyi* ADP1 ont été trouvés dans des groupes de synténie avec les 145

génomes bactériens comparés. Cette forte conservation de structure a permis de considérer l'environnement génique lors de l'annotation fonctionnelle, ce qui s'est avéré particulièrement utile avec certains gènes présentant une faible similarité de séquences avec des gènes de fonction connue (Barbe *et al.*, 2004).

2.2.3.2. Détermination de la fonction des gènes par l'utilisation d'opérons

Près de 50 % des gènes des bactéries semblent localisés au sein d'opérons. De telles structures peuvent être mise en évidence par : (i) de faibles distances entre les différents CDS ; (ii) une organisation structurale conservée au sein d'organismes différents ; (iii) un lien biologique pour les produits des gènes organisés au sein d'un même opéron ; (iv) la conservation des motifs de régulation ; (v) une vérification expérimentale, basée notamment sur la mise en évidence de l'appartenance de l'ensemble des gènes à la même unité transcriptionnelle (Brouwer *et al.*, 2008).

La plupart des approches de prédiction d'opérons développées actuellement ne permettent de travailler qu'avec des microorganismes modèles, dont le génome est séquencé et annoté (comme *Escherichia coli* ou *Bacillus subtilis*) (Kuzniar *et al.*, 2008; Li *et al.*, 2009c). Cependant, de nouvelles méthodes sont actuellement mises en place pour la prédiction d'opérons au sein de nombreux procaryotes, dont les informations génomiques ne sont que partiellement déterminées, et ce afin d'améliorer l'annotation fonctionnelle de leurs gènes. Cette approche est notamment implémentée dans UNIPOP. Cette application exploite les informations de séquences de 365 génomes bactériens, et celles de structures en opérons déjà déterminées. UNIPOP est ainsi capable de prédire, à partir de l'identification de clusters de gènes conservés au sein de génomes référents, des structures en opérons pour de nouvelles séquences, orientant ainsi l'annotation fonctionnelle des gènes de ces opérons (Li *et al.*, 2009c). De plus, les gènes organisés au sein d'un opéron, sont co-transcrits et leurs produits généralement intégrés au sein d'un même processus biologique. Ainsi, des gènes codant des produits inconnus au sein de telles structures pourraient être impliqués, par exemple, dans une même voie métabolique (Brouwer *et al.*, 2008). Si l'on connaît alors la fonction des protéines codées par les autres gènes de l'opéron, il est possible d'orienter l'annotation de ces gènes inconnus. Les structures en opérons ont, par exemple, été utilisées pour orienter l'annotation fonctionnelle de certains gènes du génome de la souche *Acinetobacter baylyi* ADP1 (Barbe *et al.*, 2004).

Tableau 7 : Principales bases de données utilisées durant l'annotation fonctionnelle de séquences et la reconstruction de voies métaboliques.

Type de base	Nom de la base	Description des ressources disponibles au sein de la base de données	Référence
Bases de données nucléiques ou d'EST	GenBank	Séquences nucléiques dites généralistes	(Benson *et al.*, 2009)
	EMBL	Séquences nucléiques dites généralistes	(Kulikova *et al.*, 2007)
	DDBJ	Séquences nucléiques dites généralistes	(Sugawara *et al.*, 2008)
	UniGene	Séquences d'EST non redondants regroupées par gènes, liés aux données d'expression et de fonction des produits des gènes	(Wheeler *et al.*, 2008)
	dbEST	Séquences d'EST disponibles	(Boguski *et al.*, 1993)
Bases de données de séquences protéiques	Swiss-Prot	Séquences protéiques de haute qualité, manuellement vérifiées	(The UniProt, 2010)
	TrEMBL	Séquences protéiques issues de la traduction des séquences nucléiques de la base EMBL	(The UniProt, 2010)
	GenPept	Séquences protéiques automatiquement extraites des CDS de la base de données GenBank	(Cochrane et Galperin, 2010)
	PIR	Séquences protéiques regroupant les données Uniprot et UniParc	(Wu *et al.*, 2003)
	COG	Séquences protéiques issus de génomes complets regroupés en familles d'orthologues	(Makarova *et al.*, 2007)
	ProtoNet	Séquences protéiques issues de Swiss-Prot regroupées en différents groupes distincts par clustering hiérarchique	(Kaplan *et al.*, 2005)
	OrthoDB	Classification de séquences protéiques de métazoaires en groupes d'orthologues	(Kriventseva *et al.*, 2008)
	BRENDA	Séquences d'enzymes ainsi que leurs propriétés : structure, spécificité, stabilité, paramètres de réactions, etc.…	(Schomburg *et al.*, 2004)
	SUPERFAMILY	Structures protéiques connues regroupées en différents groupes distincts par clustering hiérarchique	(Gough *et al.*, 2001)
Bases de données de domaines, de motifs ou de familles protéiques	Blocks	Alignements protéiques multiples représentant des domaines protéiques conservés	(Pietrokovski *et al.*, 1996)
	CDD	Alignements protéiques multiples représentant des domaines protéiques conservés, liés à des données de séquence et de structure issus de Entrez	(Marchler-Bauer *et al.*, 2009)
	Pfam	Alignements protéiques multiples représentant des domaines protéiques conservés	(Finn *et al.*, 2010)
	SMART	Domaines protéiques de haute qualité, manuellement vérifiées	(Letunic *et al.*, 2009)
	ProDom	Domaines protéiques conservés générés automatiquement à partir des ressources Swiss-Prot et TrEMBL	(Bru *et al.*, 2005)
	PRINTS	Motifs protéiques organisés pour réaliser de l'annotation fonctionnelle de protéines inconnues	(Attwood, 2002)
	PROSITE	Motifs protéiques, liés à des informations de familles protéiques, de domaines et de sites fonctionnels.	(Sigrist *et al.*, 2010)
	TIGRFAMs	Familles protéiques manuellement vérifiées liées à d'autres ressources (alignements, annotations GO, etc.…)	(Haft *et al.*, 2003)
Bases de données de réactions biochimiques et/ou de voies métaboliques	KEGG-PATHWAY	Voies métaboliques connues et référencées	(Kanehisa *et al.*, 2008)
	EMP	Voies métaboliques connues et référencées	(Selkov *et al.*, 1996)
	MPW	Voies métaboliques connues et provenant des données de génomes des bases PUMA et WIT	(Selkov *et al.*, 1998)
	EcoCyc	Voies métaboliques, transporteurs et réseaux de régulation des gènes de la souche *Escherichia coli* K-12	(Keseler *et al.*, 2009)
	MetaCyc	Voies métaboliques décrites (métabolites et enzymes) expérimentalement de plus de 1500 organismes	(Caspi *et al.*, 2010)
	BioCyc	Voies métaboliques décrites (métabolites et enzymes) de génomes totalement séquencés	(Caspi *et al.*, 2010)
	UM-BBD	Voies métaboliques spécifiques de la biodégradation de xénobiotiques (substrats, produits, enzymes impliquées, organismes connus)	(Gao *et al.*, 2010)
	PathGuide	Voies métaboliques connues et référencées, liées à d'autres données (facteurs de transcription, interactions protéines-protéines, etc.…)	(Bader *et al.*, 2006)
	PUMA2	Voies métaboliques automatiquement reconstruites de nombreux organismes en se basant sur des données d'autres bases	(Maltsev *et al.*, 2006)
	ROSY	Voies métaboliques connues spécifiques du genre *Roseobacter*	(Pommerenke *et al.*, 2008)
	SYSTOMONAS	Voies métaboliques connues spécifiques du genre *Pseudomonas* et plus spécifiquement de la souche *P. aeruginosa*	(Choi *et al.*, 2007)
	Bionemo	Voies métaboliques spécifiques de la biodégradation de xénobiotiques vérifiées manuellement de l'UM-BBD et de MetaRouter	(Carbajosa *et al.*, 2009)
	MetaRouter	Voies métaboliques spécifiques de la bioremédiation de xénobiotiques	(Pazos *et al.*, 2005)
	SKPDB	Voies métaboliques spécifiques du shikimate regroupant 8902 enzymes différentes	(Arcuri *et al.*, 2010)
Bases de données regroupant des informations diverses	Phospho3D	Structures tridimensionnelles de protéines	(Zanzoni *et al.*, 2007)
	PANTHER	Evolution des séquences protéiques (représentées par des arbres phylogénétiques), liées aux fonctions protéiques	(Mi *et al.*, 2010)
	IUBMB Enzyme List	Classement des enzymes connues (nomenclature officielle) des six grandes classes connues	(McDonald *et al.*, 2009)
	MannDB	Données diverses protéiques (propriétés chimiques, classification, fonctions, etc…) issues d'analyses protéiques dites haut-débit	(Zhou *et al.*, 2006)
	ProRule	Données structurelles et fonctionnelles provenant de PROSITE	(Sigrist *et al.*, 2005)
	TMFunction	Données fonctionnelles de protéines membranaires vérifiées expérimentalement	(Gromiha *et al.*, 2009)
	InterDom	Domaines d'interaction protéiques	(Ng *et al.*, 2003)

Toutes les approches de recherche de similarités et de signatures de séquences, les méthodes phylogénétiques, la synténie et la prédiction d'opérons peuvent être utilisées pour caractériser les fonctions des produits des gènes. Cependant, les informations obtenues n'apportent qu'une vue fragmentaire des potentialités métaboliques des organismes étudiés. En effet, pour avoir une description la plus exhaustive possible de ces voies métaboliques, il est nécessaire de relier et d'organiser ces informations.

3. Caractérisation et/ou reconstruction de voies métaboliques *in silico*

La connaissance des produits des gènes permet une description des enzymes ou des régulateurs impliqués dans un processus biologique. Dans le cas particulier des voies métaboliques, il est cependant nécessaire d'organiser les étapes où ces protéines sont impliquées afin de reconstruire dans sa totalité la voie d'intérêt. Le stockage des informations métaboliques dans des banques de données généralistes, ou spécifiques d'un organisme constitue la base de toutes les approches de reconstructions métaboliques *in silico* (Bansal, 2005; Adriaens *et al.*, 2008; Pavlopoulos *et al.*, 2008; Cochrane et Galperin, 2010).

3.1. Bases de connaissances métaboliques

A l'heure actuelle, plus de 50 bases de données réunissent les réactions biochimiques et les voies métaboliques caractérisées chez les procaryotes (Cochrane et Galperin, 2010) (Tableau 7). Les informations contenues dans ces banques concernent soit la description des réactions unitaires impliquées dans ces voies (enzymes, substrats, produits, …), soit la structuration de ces réactions afin de reconstituer les voies métaboliques dans leur globalité.

3.1.1. Banques de données d'informations métaboliques générales

Parmi ces banques d'informations métaboliques certaines stockent uniquement des données généralistes (séquences, signatures protéiques, autres) et d'autres répertorient des données sur les enzymes et/ou les métabolites comme BRENDA, LIGAND ou ENZYME (Tableau 7) (Schomburg *et al.*, 2004; Cochrane et Galperin, 2010). BRENDA, par exemple, organise les informations sur les substrats et cofacteurs connus d'une réaction unitaire donnée, ainsi que les ressources liées aux enzymes catalysant cette réaction (Schomburg *et al.*, 2004). Cette base permet donc de lister pour une réaction enzymatique, les données de cette réaction (substrat(s), produit(s), cofacteur(s),…), physicochimiques (Km, activité, pH optimum,…), moléculaires (séquence et structure de l'enzyme), ou encore taxonomiques (organismes sources).

D'autres banques regroupent ces informations (enzymes et métabolites), sous la forme d'un enchaînement de réactions unitaires, permettant ainsi de décrire les voies métaboliques caractérisées chez différents organismes (Tableau 7). Les banques KEGG-PATHWAY (Kanehisa *et al.*, 2008) et MetaCyc (Caspi *et al.*, 2010) regroupent notamment des informations expertisées sur les voies métaboliques d'organismes divers. Ces bases s'étoffent très rapidement, du fait de l'acquisition des masses de données générées par le séquençage systématique, comme le montre l'augmentation de MetaCyc (Caspi *et al.*, 2010). En 2010, 1 399 voies métaboliques étaient référencées (augmentation de 43 % par rapport à 2008), pour un total de 1 795 organismes différents (augmentation de 75 % par rapport à 2008) (Caspi *et al.*, 2008; Caspi *et al.*, 2010). Ces deux banques sont actuellement intégrées dans les centres de ressources KEGG et Biocyc (Cochrane et Galperin, 2010). A travers ces centres de ressources, de nombreuses données sont accessibles. Cependant, la recherche des voies de dégradation des HAP révèle que peu d'informations métaboliques sont disponibles (fouille de données réalisée en juillet 2010). Pour ces deux bases, les voies du naphtalène sont disponibles, et seul KEGG-PATHWAY décrit les voies de dégradation du phénanthrène et de l'anthracène, bien que tous les métabolites ne soient pas présents. D'autres voies sont totalement absentes, comme par exemple les voies de dégradation du fluoranthène et du pyrène qui ne sont pas retrouvées au sein de ces bases.

3.1.2. Banques de données d'informations métaboliques ciblées

D'autres bases contiennent des informations spécifiques de voies métaboliques d'organismes, ou de groupes d'organismes modèles, ou d'intérêt médical et/ou biotechnologique (Tableau 7). Ainsi, de manière non exhaustive et pour exemple, les banques EcoCyc, SYSTOMONAS et ROSY peuvent être citées. La première est dédiée à l'espèce *Escherichia coli* et regroupe actuellement 253 voies métaboliques, 1 445 enzymes et 1 829 réactions métaboliques (incluant les réactions de transport) (Keseler *et al.*, 2009). SYSTOMONAS quant à elle, est spécifique du genre *Pseudomonas* et plus particulièrement de l'espèce pathogène *P. aeruginosa* (Choi *et al.*, 2007). Elle contient 147 voies métaboliques, et 7 567 réactions métaboliques, pour un total de 15 souches de 7 espèces de *Pseudomonas* différentes. Enfin, la banque de données ROSY (Pommerenke *et al.*, 2008) est consacrée au phylotype le plus représenté dans les écosystèmes marins, *Roseobacter*. Elle regroupe 137 voies métaboliques et 92 560 protéines, pour un total de 22 espèces à fort potentiel biotechnologique. D'autres bases vont être plus spécifiques des voies métaboliques centrées autour de précurseurs métaboliques. C'est notamment le cas de la banque de données

Tableau 8 : Principales approches dédiées à la consultation de données, à travers une interface graphique.

Nom	Description de l'approche implémentée ainsi que des avantages de l'outil	Référence
BioJAKE	Visualisation, création et manipulation de voies métaboliques sans vérification. Stockage des voies au sein d'une base de données locale dédiée simple. Aucune voie fournie au sein de la base de données initiale. La visualisation graphique se fait de manière particulière via des molécules (métabolites ou enzymes) et des flèches pour les réactions.	(Salamonsen *et al.*, 1999)
paVESY	Visualisation, création et manipulation de voies métaboliques sans vérification. Stockage des voies au sein d'une base de données locale dédiée qui est le principal intérêt de cette approche. Ainsi, les types de données conservées dans la base sont modifiables et adaptables aux besoins de chacun (données de génomique, de protéomique, etc…). Aucune voie fournie au sein de la base de données initiale.	(Ludemann *et al.*, 2004)
PathVisio	Visualisation, création et manipulation de voies métaboliques sans vérification. Stockage des voies au sein d'une base de données locale dédiée. Aucune voie fournie au sein de la base de données initiale. Possibilité de lier des données externes (liens vers des bases ou des références bibliographiques) à chaque élément de la voie métabolique.	(van Iersel *et al.*, 2008)
KGML-ED	Visualisation et manipulation de voies métaboliques sans vérification logicielle, issues de la base de données KEGG. Stockage des voies au sein d'une base de données locale dédiée au format spécifique KEGG. La visualisation graphique est basée sur celle de KEGG.	(Klukas et Schreiber, 2007)
FMM	Visualisation de voies métaboliques uniquement. Cette visualisation s'appuie sur les données de KEGG, LIGAND et UniProt pour reconstruire le passage d'un métabolite à un autre via plusieurs voies métaboliques différentes. Pas de possibilité d'ajout manuel de données au sein de la base statique.	(Chou *et al.*, 2009)
SubPathwayMiner	Visualisation des voies métaboliques basée sur les données des voies de KEGG, par une recherche d'identifiants de gènes ou d'enzymes. Les données sont stockées dans une base locale, les données KEGG sont directement importables. La mise à jour est automatique. La recherche et la visualisation des voies sont simplifiées (une réaction enzymatique est considérée comme un seul élément graphique).	(Li *et al.*, 2009b)
ReactionExplorer	Visualisation de voies métaboliques par recherche de synonymes au sein de la base IUBMB Enzyme List pour chaque enzyme ou protéine impliquée. Stockage des données au sein d'une base de données locale basée sur l'IUBMB Enzyme List.	(McDonald *et al.*, 2009)
PathCase	Visualisation de voies métaboliques qui s'appuie sur les données de KEGG, de la littérature et des données BioCyc d'*Homo sapiens*. Stockage des données au sein d'une base de données accessible via le Web. La visualisation graphique se fait de manière classique. La recherche et la sélection peut se faire à de nombreux niveaux : nom de métabolite, enzyme, voie métabolique, organismes, etc…	(Elliott *et al.*, 2008)
Cyclone	Visualisation de voies métaboliques via un système informatique plus souple que celui existant pour analyser et éditer les données de BioCyc. Stockage des données au sein d'une base de données locale. La visualisation graphique se fait de manière classique. L'intérêt principal est que, via cette approche, il est possible d'enrichir facilement la base BioCyc en fournissant de nouvelles voies déjà structurées pour la banque BioCyc.	(Le Fèvre *et al.*, 2007)

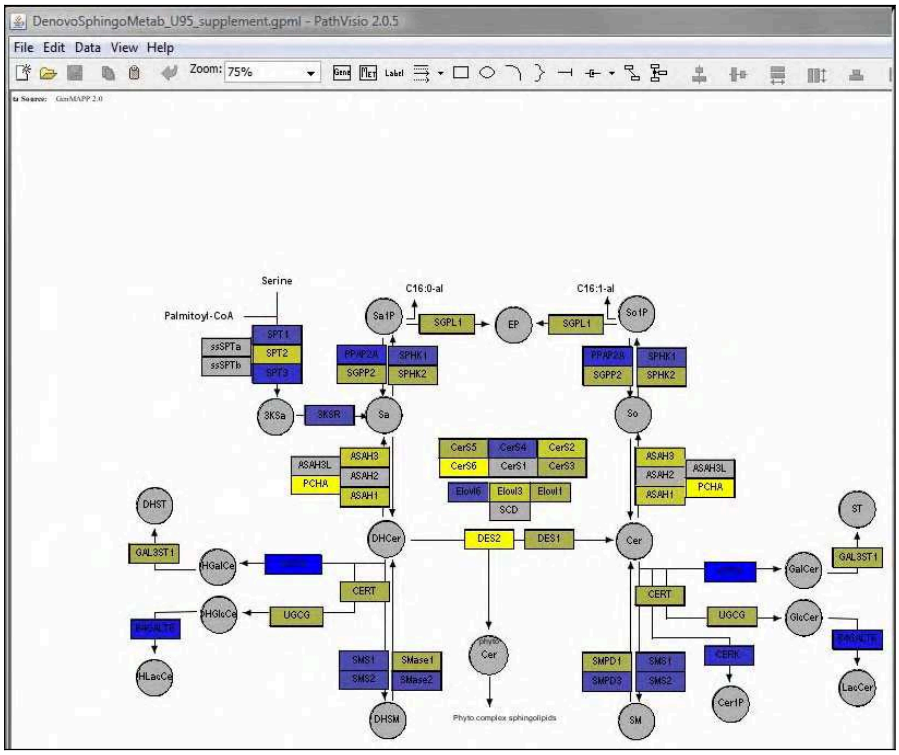
SKPDB (ShiKimate Pathway DataBase), qui répertorie les voies ou les réactions impliquant le shikimate, précurseur des acides aminés aromatiques ou de l'acide salicylique. Cette base regroupe 8 902 enzymes différentes (Arcuri *et al.*, 2010). D'autres bases vont regrouper, quant à elles, toutes les voies métaboliques impliquées dans la biodégradation de polluants comme les banques MetaRouter (Pazos *et al.*, 2005) ou encore l'UM-BBD (University of Minnesota Biocatalysis/Biodegradation Database) (Gao *et al.*, 2010) (Tableau 7). L'UM-BBD est mise à jour mensuellement et contient actuellement 191 voies, 1 306 réactions, 1 213 composés, 851 enzymes et 497 microorganismes différents. Cette banque de données, d'un fort intérêt biologique pour notre étude, est donc très riche en informations sur la dégradation des polluants, qu'ils soient aromatiques ou non. Cependant, elle ne permet pas de retrouver facilement tous les organismes capables de réaliser la voie d'intérêt (ni les informations sur l'organisation des opérons codant pour les enzymes d'intérêt). De même, les séquences protéiques recherchées ne peuvent être récupérées qu'à travers UNIPROT, enzyme après enzyme.

De plus, la spécificité propre à chacune de ces banques de données de connaissances métaboliques a entraîné une forte hétérogénéité de leur structuration, mais également une forte hétérogénéité des formats de stockage des informations. De plus, les méthodes de consultations statiques de ces banques limitent généralement l'intégration des données personnelles, ou de croiser les informations entre plusieurs sources. Ces contraintes, combinées à l'augmentation régulière de la masse des données, empêchent d'exploiter simplement et efficacement ces informations pour consulter, et/ou reconstruire des voies métaboliques d'intérêt (Adriaens *et al.*, 2008). De nombreuses approches bioinformatiques sont actuellement développées (outils de consultation, d'annotations, de reconstruction métabolique, de fouille de données, etc.…) pour s'affranchir de ces limitations, et faciliter ainsi l'exploitation exhaustive de ces données (Bansal, 2005; Adriaens *et al.*, 2008; Karimpour-Fard *et al.*, 2008; Pavlopoulos *et al.*, 2008). Ces approches seront présentées et décrites au sein du paragraphe suivant ainsi que les outils informatiques qui en découlent.

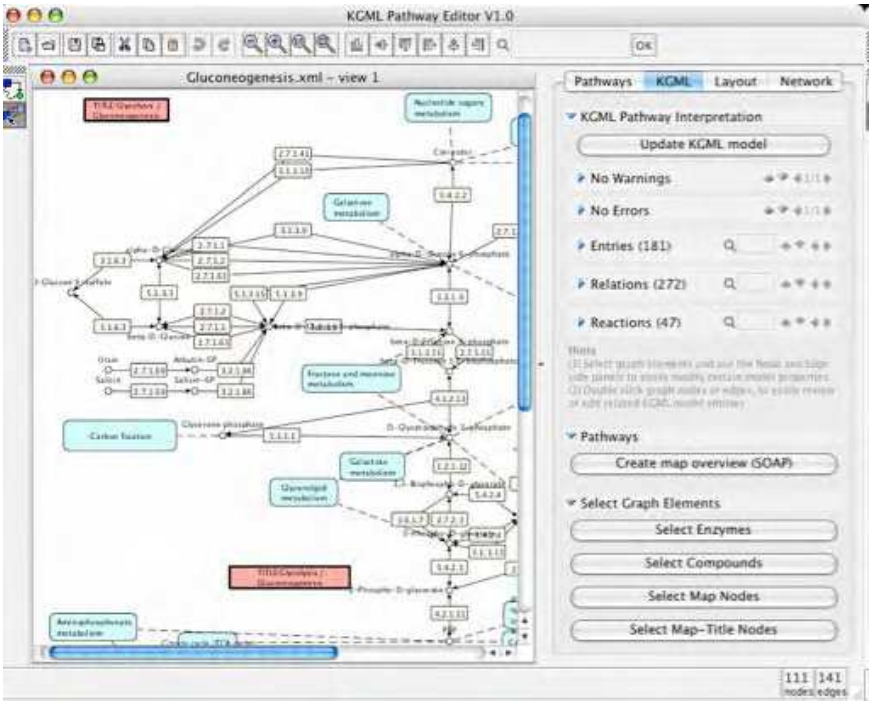
3.2. Outils pour la consultation ou la fouille de données métaboliques

3.2.1. Outils de consultation

De nouvelles approches ont été développées pour favoriser une consultation (recherche, visualisation, édition) plus dynamique des ressources disponibles dans les banques (Tableau 8). Ces outils (sauf BioJAKE) ne permettent cependant pas d'inclure des données



<u>Figure 22 :</u> Capture d'écran d'une voie métabolique visualisée avec PathVisio.
Expression des gènes codant pour les protéines impliquées dans une voie métabolique de production de sphingolipides obtenues par une analyse biopuce ADN. Les métabolites sont représentés par des ronds gris, les gènes par des rectangles. La surexpression des gènes est visible en jaune, et la sous-expression en bleu selon les deux conditions testées. Avec les options de la barre d'outils, plusieurs éditions sont possibles (ajout d'arcs, de gènes de données, etc.) (van Iersel *et al.*, 2008).



<u>Figure 23 :</u> Capture d'écran d'une voie métabolique visualisée avec KGML-ED.
Cet outil permet de visualiser au format graphique KEGG une vois métabolique définie au sein de la base KEGG. Un autre avantage est la possibilité d'éditer cette voie (suppression, déplacement, ajout de couleurs, changement de noms, etc.…).

totalement personnelles et d'effectuer de la fouille de données en tant que telle (à par avec des mots-clés). Ils sont uniquement des outils de visualisation de voies métaboliques.

Certaines de ces approches ne sont dédiées qu'à la visualisation des informations métaboliques, sans contrôle de la qualité des données, puisqu'elles laissent le libre choix à l'utilisateur d'intégrer ses informations personnelles pour assurer la reconstruction des voies métaboliques. Parmi les applications implémentant ces approches (Tableau 8), certaines comme paVESY (Ludemann *et al.*, 2004), ou PathVisio (van Iersel *et al.*, 2008) permettent de relier les voies construites à d'autres données. Ainsi, l'utilisateur a la possibilité de rajouter sur les voies métaboliques reconstruites des données hétérogènes numériques (PCR quantitative, transcriptomique, etc.), ou textuelles (annotation fonctionnelle, lien hypertexte vers d'autres banques, GO, etc.) apparaissant sous forme de couleurs ou de liens (Figure 22). Un des principaux avantages de PathVisio par rapport à paVESY, est que ces données peuvent être importées sous différentes formes : via des fichiers texte (avec des séparateurs de type tabulations, des fichiers Excel, des fichiers .csv ou encore manuellement.

D'autres approches de consultation donnent quant à elles, la possibilité d'éditer des cartes métaboliques contrôlées puisqu'elles s'appuient uniquement sur les données expertisées de KEGG. Certaines de ces approches associent également des données hétérogènes aux cartes métaboliques d'intérêt en utilisant un système de couleurs, représentant des niveaux d'expression par exemple (PCR quantitative, transcriptomique, protéomique, etc.). PathwayExplorer (Mlecnik *et al.*, 2005), KEGGanim (Adler *et al.*, 2008) et PathExpress (Goffard et Weiller, 2007) ont été développées dans ce sens. D'autres approches proposent une modification de la structure des cartes métaboliques issues de KEGG, pour une exploration plus dynamique de ces données. Ainsi, il est possible, avec des applications comme KGML-ED (Klukas et Schreiber, 2007), FMM (Chou *et al.*, 2009), SubPathwayMiner (Li *et al.*, 2009b), et PathCase (Elliott *et al.*, 2008) de visualiser et d'éditer ces informations en local (Figures 22 et 23) (Tableau 8). Les ressources propres à BioCyc (incluant entre autre MetaCyc et EcoCyc) peuvent, quant à elles, être visualisées et analysées à l'aide de l'outil Cyclone qui leur est spécifiquement dédié (Le Fèvre *et al.*, 2007). Tous ces outils ont cependant comme principale limite qu'il est impossible d'inclure des données personnelles au sein des voies, mais uniquement d'éditer les voies déjà définies au sein des bases KEGG et BioCyc (il est ainsi possible de créer de nouvelles étapes enzymatiques, mais sans vérification logicielle). De plus, ces outils restent des outils de consultation, la fouille des données (à part la recherche des voies visualisées) n'étant pas implémentée.

Des approches de consultation plus complexes peuvent assurer la comparaison des voies métaboliques présentes chez plusieurs organismes. Il est ainsi possible d'identifier les différences ou les similitudes métaboliques entre deux espèces en se basant uniquement sur la structure des cartes métaboliques. Le logiciel FIT-MATCH exploite ainsi les données d'annotation de KEGG (métabolite, enzymes, identifiants, etc.…) pour réaliser ce type d'étude comparative (Wernicke et Rasche, 2007). FIT-MATCH possède cependant une limite majeure : il est impossible avec cet outil d'utiliser des données personnelles, seules les données définies de KEGG sont utilisables, contrairement à d'autres outils comme METAPAT ou Comparative Pathway Analyzer (Wernicke et Rasche, 2007; Oehm *et al.*, 2008).

Les informations métaboliques regroupées dans les banques, comme KEGG et BioCyc, ne sont cependant pas compatibles entre elles, car elles présentent des formats de stockage différents. Il est donc difficile de les exploiter de façon exhaustive et de croiser leurs données respectives. Des approches d'agrégation de ces données métaboliques sont donc proposées de façon à mutualiser ces données. L'outil cPATH intègre ainsi les informations provenant de plusieurs sources (MINT, IntAct, HPRD, DIP, BioCyc, KEGG, PUMA2 et Reactome) permettant le stockage, la visualisation et l'analyse de l'ensemble des données métaboliques de ces différentes bases (Cerami *et al.*, 2006).

Ces approches de consultations des informations métaboliques favorisent donc la recherche, la visualisation, voire l'édition des données hétérogènes des cartes métaboliques, issues ou non de banques de données expertisées. Cependant, il est impossible, via les applications décrites dans ce paragraphe: (i) de reconstruire *in silico* de nouvelles voies métaboliques à partir de données d'annotation ou de séquence ; (ii) de comparer des données personnelles avec celles disponibles dans les banques.

3.2.2. Outils de fouille des données et de reconstruction

3.2.2.1. Fouille de données et reconstruction *in silico* de voies métaboliques

Les approches de prédiction des fonctions des produits des gènes (similarité de séquence, recherche de signatures, méthodes phylogénétiques, synténie, prédiction d'opérons) permettent d'attribuer à des séquences inconnues des mots-clés, définissant leur fonction. Ces descripteurs s'appuient sur un vocabulaire contrôlé permettant de décrire les propriétés des réactions biochimiques (nom systématique de l'enzyme, numéro EC, etc.) des réactions métaboliques des organismes étudiés. Pour reconstruire les cartes métaboliques d'organismes d'intérêt, il est cependant nécessaire de caractériser et de réorganiser les relations entre ces

Tableau 9 : Principales approches dédiées à la fouille des données et à la reconstruction métabolique.

Catégorie	Nom	Description de l'approche implémentée, des données utilisées en entrée et des résultats fournis par l'outil informatique implémenté.	Référence
Approches se basant sur les informations d'annotation disponibles et sur celles regroupées au sein de bases comme KEGG ou MetaCyc	PathFinder	La reconstruction dynamique de voies métaboliques est basée sur l'utilisation d'un génome annoté. En croisant les données disponibles au sein de KEGG et les données d'annotation du génome analysé (numéros EC, noms, etc…), PathFinder reconstruit les voies métaboliques de manière dynamique. Il est également possible de les éditer et de les sauvegarder en local.	(Goesmann *et al.*, 2002)
	PathMiner	L'idée développée permet de reconstruire de manière automatique les voies métaboliques d'un génome annoté avec des données biochimiques. La reconstruction et la recherche se base sur les données disponibles de KEGG, et non sur la similarité de séquence et sur les données biochimiques fournies. Les résultats sont représentés via une interface graphique proche de celle de KEGG.	(McShan *et al.*, 2003)
	Pathway Tool Software	L'intérêt repose ici sur la possibilité de reconstruire les voies métaboliques d'un organisme dont le génome a été séquencé. En utilisant les données d'annotation correspondantes (numéro EC des enzymes, substrats, positions et nom du gène, etc…), il est possible de reconstruire les voies impliquées en corrélant ces données à celles disponibles dans la base MetaCyc.	(Karp *et al.*, 2002)
	MinPath	Son originalité repose d'abord sur la possibilité d'utiliser des données d'annotation de génomes ou de métagénomes. Ainsi, à partir des fonctions de chacune des protéines fournies par l'utilisateur, l'approche s'appuie sur la minimisation du nombre de voies pour toutes ces fonctions. Il reconstruit ainsi un certain nombre de voies en incluant les fonctions fournies en se basant sur les voies de KEGG.	(Ye et Doak, 2009)
Approches se basant sur la similarité de séquences protéiques et sur les données de voies disponibles au sein de bases comme KEGG, EMP ou MetaCyc	WIT	L'approche consiste à reconstruire les voies métaboliques de génomes séquencés en se basant sur les données de 40 génomes déjà séquencés et dont les voies ont été définies dans les bases EMP et MPW. Cette détermination de voies s'appuie sur la similarité de séquence mais aussi sur la synténie entre génomes.	(Overbeek *et al.*, 2000)
	PATH-A	L'approche consiste à reconstruire les voies métaboliques à partir des séquences protéiques choisies en se basant sur des voies prédéfinies. La comparaison de séquence (BLAST, HMM, etc...) permet de définir les enzymes impliquées dans la voie choisie. Actuellement, 10 voies métaboliques ont été développées et sont disponibles. Une représentation graphique est également fournie à travers le logiciel implémenté.	(Pireddu *et al.*, 2006)
	Pathway Voyager	L'approche mise en place ici permet de reconstruire les voies métaboliques de procaryotes en se basant sur les séquences protéiques fournies par l'utilisateur. Ces dernières sont comparées par BLASTp à celles de la base de données KEGG disponibles, afin de replacer chacune des protéines au sein d'une ou plusieurs voies métaboliques. Cet outil permet d'outrepasser les problèmes d'annotation en s'appuyant sur la similarité de séquence.	(Altermann et Klaenhammer, 2005)
	FUNGIpath	L'approche est axée sur la prédiction d'orthologues à partir des séquences protéiques fournies par l'utilisateur. Cette étape est réalisée avec une base de génomes de champignons. Il va donc définir la fonction de chacune des protéines. Puis, en se basant sur KEGG et MetaCyc, il reconstruit les voies métaboliques présentes en fonction des enzymes retrouvées.	(Grossetete *et al.*, 2010)
Approches se basant sur la comparaison de voies métaboliques entre organismes	Comparative Pathway Analyzer	L'approche mise en place consiste en la comparaison de la même voie métabolique de deux organismes différents disponibles au sein de la liste de voies disponibles. Les enzymes impliquées sont comparées par orthologie et les voies métaboliques proviennent de KEGG ou de données personnelles si elles sont au format décrit.	(Oehm *et al.*, 2008)
	METAPAT	L'approche consiste en la comparaison de deux voies métaboliques à partir des numéros EC (ou numéros Enzyme Commission) d'organismes différents. Il est possible d'utiliser au sein du logiciel implémenté des données provenant de BioCyc ou des voies personnelles au même format. Les différences sont mises en valeur ainsi que les points communs via l'interface graphique.	(Wernicke et Rasche, 2007)

réactions métaboliques unitaires. Pour cela, deux approches peuvent être envisagées, implémentées au sein de divers outils plus ou moins spécifiques.

La première utilise uniquement les descripteurs en s'appuyant sur les banques de données métaboliques généralistes (Tableau 9). En effet, ces bases de connaissances (comme KEGG ou MetaCyc) regroupent des cartes métaboliques expertisées, qui peuvent permettre de déterminer le potentiel métabolique de l'organisme étudié. Pour cela, il est notamment possible d'interroger ces banques, en leur fournissant des descripteurs (numéro EC par exemple) pour les protéines étudiées de l'organisme d'intérêt. Les voies métaboliques faisant intervenir ces protéines chez d'autres organismes, sont alors utilisées pour reconstruire les voies potentielles. Les outils informatiques comme PathFinder (Goesmann *et al.*, 2002), PathMiner (McShan *et al.*, 2003), MinPath (Ye et Doak, 2009) ou le 'Pathway Tool Software' (Karp *et al.*, 2002) ont été implémentés dans ce but. L'originalité de l'application MinPath repose sur la minimisation du nombre de voies métaboliques reconstruites en fonction du nombre de protéines impliquées. Ceci permet ainsi de limiter les redondances si des protéines sont impliquées dans de nombreuses voies, car seule la voie faisant intervenir un maximum de protéines, également recherchées, sera reconstruite (Ye et Doak, 2009). Cependant, ces outils se basant sur les données disponibles au sein des bases KEGG et MetaCyc, il est difficile de reconstruire les voies impliquées dans la dégradation des HAP. En effet, comme précisé précédemment, peu de données sur ces voies de dégradation sont référencées et décrites au sein de ces bases.

La seconde approche réalise directement l'annotation fonctionnelle des produits des gènes étudiés, par recherche de similarité, puis réorganise ces différents constituants métaboliques pour reconstruire des voies. La recherche de similarité de séquences s'effectue grâce à l'algorithme BLAST, ou par des approches phylogénétiques (Altermann et Klaenhammer, 2005; Pireddu *et al.*, 2006; Grossetete *et al.*, 2010) contre des banques de données comme KEGG ou BioCyc (Tableau 9). Cette méthode est implémentée dans des outils comme WIT (Overbeek *et al.*, 2000), Pathway Voyager (Altermann et Klaenhammer, 2005), FUNGIpath (Figure 24 pages suivante) (Grossetete *et al.*, 2010), ou PATH-A (Pireddu *et al.*, 2006), qui sont fortement adaptés à l'analyse de génomes complets. Il est cependant possible de ne rechercher qu'une voie métabolique particulière dans les bases, si celle-ci est bien sûr référencée.

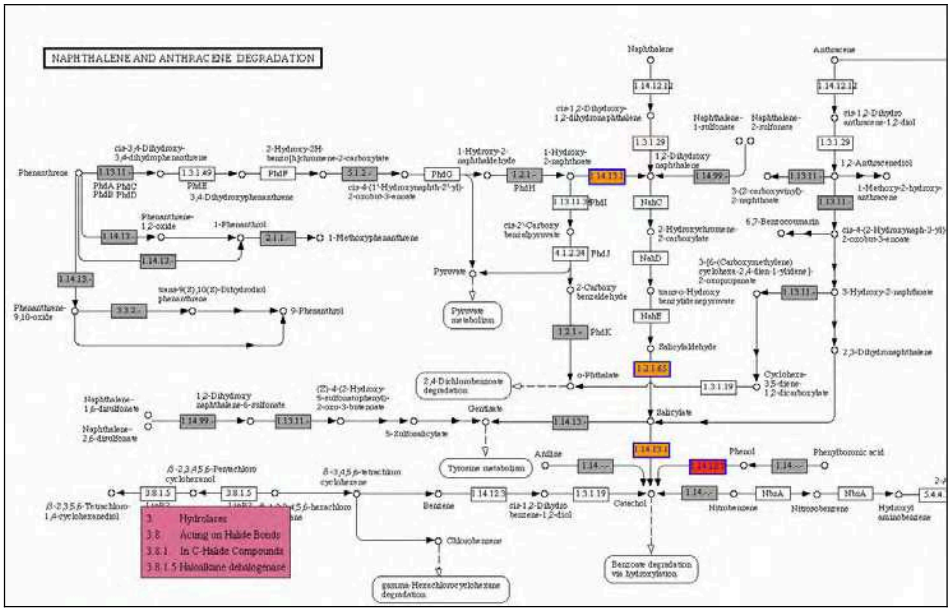


Figure 24 : Capture d'écran d'une voie métabolique partielle visualisée avec FungiPATH.
Voie métabolique décrite au sein de la banque KEGG pour la dégradation des HAP (naphtalène, phénanthrène et anthracène) pour tous les organismes de la base FungiPATH. Les enzymes sont représentés par des numéros EC au sein de cases rectangulaires, les métabolites par des ronds blancs. Les couleurs (allant du blanc au bordeaux) indiquent la présence (en pourcentage de génomes analysés) contenant l'enzyme recherchée. En gris sont les enzymes non définies précisément, où la recherche n'est pas effectuée. Pour chaque enzyme, il est également possible d'afficher une fenêtre interne donnant plus de détails sur la protéine considérée (affichée en rouge) (Grossetete *et al.*, 2010).

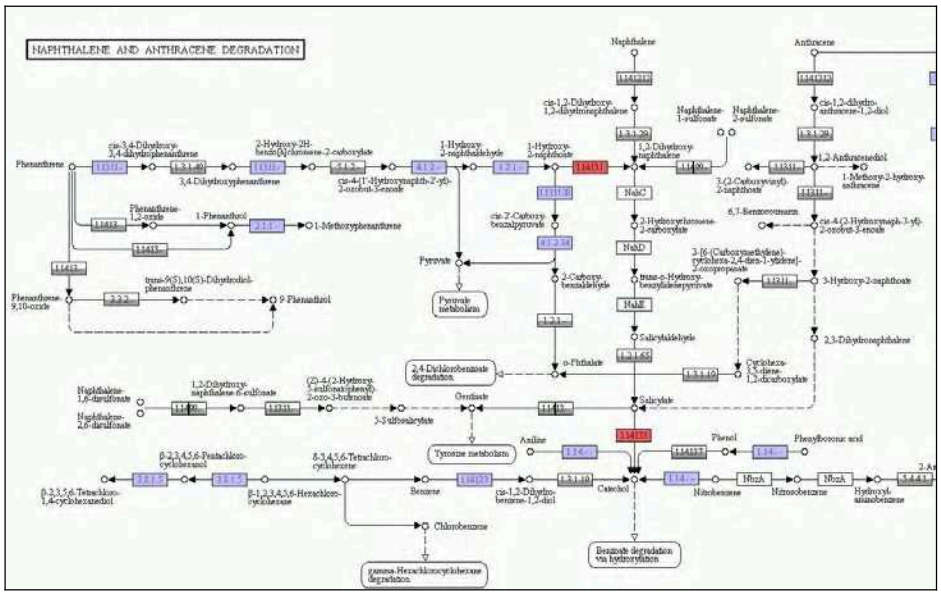


Figure 25 : Capture d'écran d'une voie métabolique visualisée avec Comparative Pathway Analyzer.
Voie métabolique décrite au sein de la banque KEGG pour la dégradation des HAP (naphtalène, phénanthrène et anthracène) pour les organismes *Pseudomonas putida* F1, *Mycobacterium vanbaalenii*, *Mycobacterium* sp. JLS et *Sphingomonas wittichii* (Oehm *et al.*, 2008). Les enzymes sont représentées par des numéros EC au sein de cases rectangulaires, les métabolites par des ronds blancs. En rouge apparaissent les enzymes retrouvées chez *Pseudomonas putida* F1 uniquement, en gris les enzymes non retrouvées, en bleu les enzymes retrouvées chez au moins un des deux espèces de *Mycobacterium*.

3.2.2.2. Comparaison de voies métaboliques existantes

Il est également intéressant de pouvoir facilement comparer les potentialités métaboliques entre organismes, dans le but, par exemple de pouvoir caractériser les relations métaboliques qui existent entre ces derniers, comme celles au sein de *consortia* (Tableau 9) (Wernicke et Rasche, 2007; Oehm *et al.*, 2008). Pour cela, certaines approches utilisent les mots-clés provenant des annotations fonctionnelles, d'autres les informations issues de la recherche de similarité de séquences. Le logiciel METAPAT exploite ainsi uniquement les numéros EC, et les noms systématiques des métabolites des deux voies comparées (Wernicke et Rasche, 2007). Cela permet, pour une voie métabolique de deux organismes différents, de montrer les similitudes et les différences pour ces deux organismes. Cependant, cette comparaison nécessite donc des voies parfaitement annotées et vérifiées. C'est pourquoi, ces voies métaboliques proviennent généralement de la base BioCyc, mais peuvent aussi provenir de données personnelles.

Les comparaisons effectuées sont plus rapides qu'avec les données de séquences, car seuls les numéros EC et les noms sont utilisés, mais dépendent de la qualité des annotations fonctionnelles. Comparative Pathway Analyzer (Figure 25) s'appuie, quant à lui, sur les données d'orthologie (similarité de séquences) pour comparer les produits de gènes impliquées dans les voies métaboliques étudiées (Oehm *et al.*, 2008). Les voies métaboliques qui peuvent être comparées via cet outil s'appuient sur les données de KEGG de 155 eucaryotes, de 569 bactéries et de 49 archées.

4. Conclusion

L'annotation fonctionnelle reste la première étape systématique de la reconstruction *in silico* de voies métaboliques, mais les informations obtenues n'apportent qu'une vue isolée des potentialités métaboliques des organismes étudiés. En effet il est nécessaire de réorganiser ces informations métaboliques *in silico* pour avoir un réel aperçu des capacités métaboliques de l'organisme d'intérêt.

Néanmoins, ces analyses doivent être vérifiées expérimentalement, afin de confirmer les capacités métaboliques des organismes étudiés. Cette vérification, réalisée jusqu'à récemment par des approches ciblées (expression hétérologue, mutagénèse dirigée, etc.…), n'est plus possible sur l'ensemble des données générées par les approches de génomiques. Il est cependant nécessaire de pouvoir exploiter avec pertinence l'ensemble de ces séquences pour caractériser les voies métaboliques, mais également leurs régulations et les relations

métaboliques qui peuvent exister entre les différents organismes. L'exploitation de la richesse de ces données de génomique peut être entreprise pour le développement d'outils de post-génomique utiles pour l'exploration d'échantillons biologiques plus ou moins complexes. Dans ce cadre, les biopuces ADN représentent une approche de choix permettant d'identifier simultanément plusieurs centaines de milliers de gènes avec la possibilité de traiter en une seule expérience un grand nombre d'échantillons. Nous présenterons dans le paragraphe suivant les approches biopuces ADN, leurs conceptions utilisant les données de génomique et leurs applications en écologie microbienne.

Chapitre IV : Biopuces ADN pour l'étude des capacités métaboliques des microorganismes

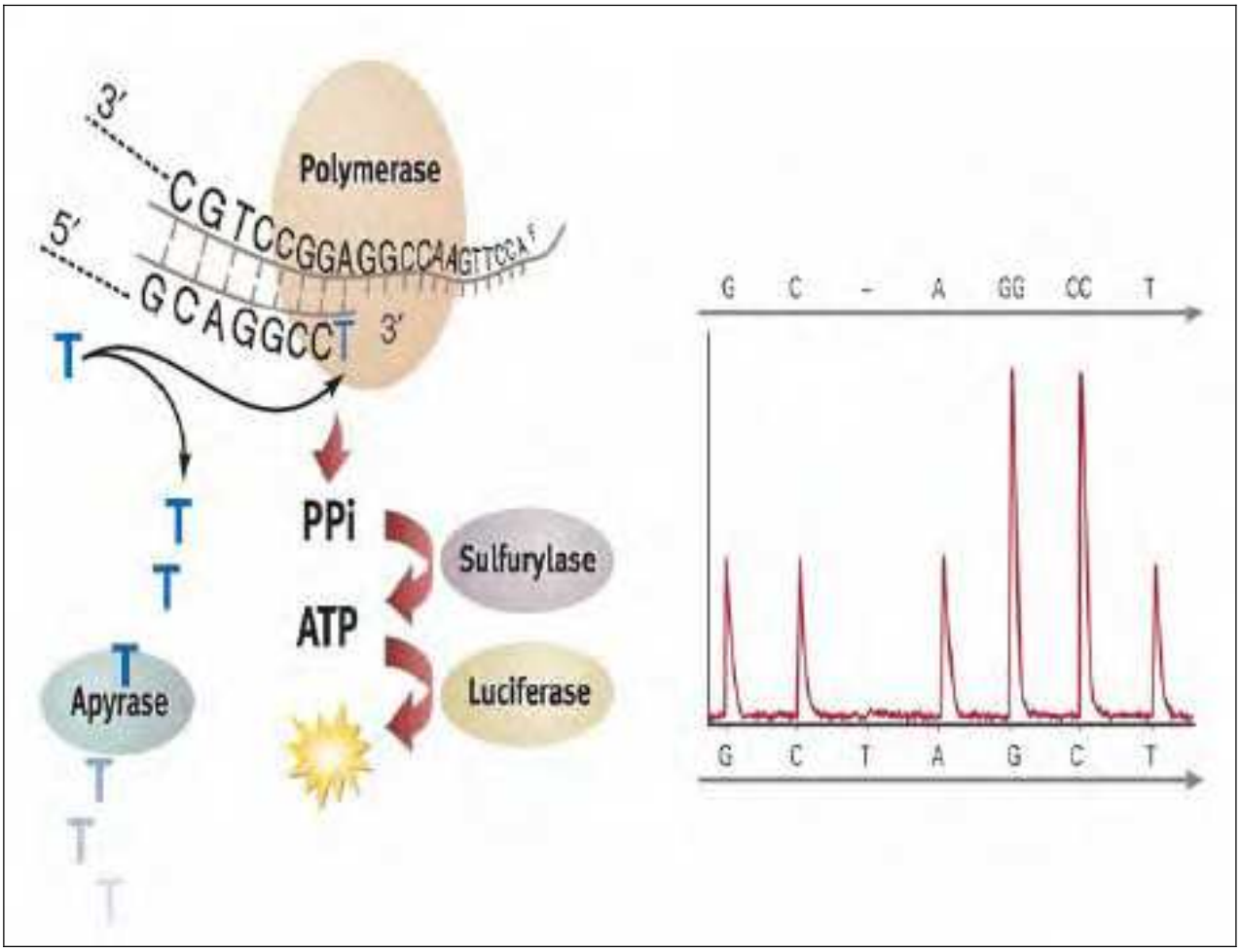
1. Introduction

Depuis le développement de l'ancêtre du microscope, et les observations par Antonie van Leeuwenhoek en 1676, de nombreux chercheurs tendent à développer des techniques permettant de cultiver et d'identifier les microorganismes. Un des premiers fut Louis Pasteur qui, en 1861, développa les techniques de pasteurisation et de stérilisation qui permettront dès lors d'isoler les microorganismes et de les étudier en cultures pures. A la fin du XIXème siècle, J.R. Petri inventa la boîte de Petri, permettant la culture en conditions stériles des microorganismes.

Néanmoins, depuis ces dernières années, l'écologie microbienne se base de plus en plus sur l'étude des écosystèmes dans leur globalité, ceci afin d'acquérir de nouvelles connaissances sur les interactions entre les microorganismes (comme les *consortia* qui ont notamment des capacités métaboliques complémentaires). En effet, la grande majorité de ces microorganismes provenant d'écosystèmes complexes ne peut être cultivée avec les techniques existantes à ce jour. Pour appréhender cette diversité dans sa globalité, de nouvelles approches moléculaires sont donc utilisées (amplification génique, empreintes génétiques, séquençage, hybridation *in situ*, biopuces ADN). La caractérisation d'environnements complexes nécessite donc l'emploi d'outils à haut débit pour obtenir une résolution suffisante assurant une bonne exploration du modèle étudié.

2. La révolution moléculaire en écologie microbienne, les nouveaux outils haut-débit

Jusqu'à récemment, la mise en culture était une étape obligatoire dans l'identification des microorganismes et la caractérisation de leurs capacités métaboliques. Or, ces techniques de culture par enrichissement, utilisées classiquement, ne permettent d'isoler qu'une très faible fraction des populations microbiennes (Leadbetter, 2003). En effet, même si de nouvelles techniques de culture, utilisant des milieux pauvres en nutriments avec une grande diversité de substrats (Ferrari *et al.*, 2005), représentent un réel potentiel et une avancée considérable, elles s'avèrent encore inadaptées pour appréhender la totalité de la biodiversité des écosystèmes. Actuellement, les études réalisées estiment que moins de 1 % des 10^9



<u>Figure 26 :</u> Principe du pyroséquençage.
Contrairement au séquençage classique, les nucléotides ne sont pas ajoutés tous ensemble mais les uns après les autres. Si le nucléotide ajouté est complémentaire de la matrice, l'ADN polymérase l'incorporera en libérant un pyrophosphate (PPi). Ce dernier va être transformé en ATP grâce à une ATP sulfurylase. L'ATP ainsi formé sera alors utilisée par une luciférase pour émettre un signal lumineux. Ce signal va être capté par le séquenceur et traduit en un pic qui sera plus ou moins important en fonction du nombre de nucléotides incorporés. Les nucléotides qui n'ont pas été incorporés par la polymérase sont dégradés par une apyrase avant qu'un nouveau nucléotide ne soit ajouté.
Source (Invitrogen) : http://www.invitrogen.com/site/us/en/home/Global/molecular-biology-cell-biology-applications.html

bactéries présentes dans un gramme de sol sont cultivées en laboratoire (Davis *et al.*, 2005). La complexité des interactions métaboliques microbiennes est donc très difficile à reproduire avec les techniques d'isolement existantes actuellement (Saleh-Lakha *et al.*, 2005).

Les nouvelles stratégies de caractérisation des structures des communautés microbiennes sont de plus en plus basées sur l'utilisation de techniques de biologie moléculaire, s'appuyant sur l'ADN et/ou l'ARN. En effet, l'extraction du matériel génétique qui peut être réalisée directement à partir d'un échantillon environnemental, sans étape culturale intermédiaire facilite encore l'étude directe des écosystèmes (Galvão *et al.*, 2005). De plus, les progrès et les innovations du séquençage ont permis l'amélioration des techniques existantes, mais aussi le développement de nouvelles techniques, comme le pyroséquençage (Figure 26) (Ronaghi, 2001; Margulies *et al.*, 2005; Hall, 2007). Cette technique dite de très-haut-débit, permet le séquençage de plus de 30 Mpb, avec une longueur moyenne de séquences de 400 à 500pb.

Ces nouvelles capacités de séquençage ont facilité l'apparition et la démocratisation de nouvelles méthodes haut-débit dites de méta«omiques» (principalement la métagénomique et la métatranscriptomique), ainsi que les biopuces ADN (Stenuit *et al.*, 2008). La métagénomique consiste en l'étude de l'ensemble des génomes contenus au niveau d'un environnement, par un séquençage systématique. Après séquençage, les fragments de génomes sont assemblés et assignés à une souche grâce à la présence potentielle de biomarqueurs, mais également par l'estimation du pourcentage en GC, de fréquences des di-, tri- et tétranucléotides du biais d'usage des codons de chaque séquence. Outre le fait d'identifier de nouvelles espèces microbiennes, la métagénomique permet l'étude de l'immense réservoir génétique microbien des écosystèmes et plus particulièrement de sa fraction non cultivée (Schmeisser *et al.*, 2007). Une étude récente des communautés microbiennes d'un écosystème aquatique contaminé par des métaux lourds (principalement de l'uranium) et des solvants organiques, a permis l'obtention d'environ 53Mpb de données de bonne qualité (sur 78Mbp de données brutes) (Hemme *et al.*, 2010). Cette étude a permis de mettre en évidence les adaptations des communautés microbiennes à des pollutions présentes depuis plus de 50 ans. En effet, les communautés présentes montrent une très faible diversité, avec une surabondance de la présence des gènes de résistance aux polluants présents (lié à de nombreux transferts latéraux de gènes).

A l'instar de la métagénomique, des études visant à identifier l'ensemble des activités exprimées *in situ* par les microorganismes voient le jour. Cette approche originale, nommée métatranscriptomique, est actuellement encore peu utilisée sur les communautés microbiennes

de sols. Son utilisation repose sur l'extraction des ARN totaux et sur le clonage des produits obtenus après transcription inverse. Une étude originale, publiée en 2009, compare les métatranscriptomes d'un même écosystème aquatique (mer hawaïenne) exposé à la lumière du jour ou durant la nuit (Poretsky *et al.*, 2009). Cette étude a permis d'obtenir 75 558 séquences pour le jour et 75 946 séquences pour la nuit, l'analyse des données révélant de nombreuses différences d'expression de gènes en réponse à l'ensoleillement. Ces techniques, bien que donnant de nombreuses informations de séquence, nécessitent un équipement important, et ne donnent un aperçu qu'à un moment précis de l'écosystème étudié. De plus, ces nouvelles méthodes nécessitent le développement d'outils informatiques spécifiques pour le traitement et l'analyse des données. De même, l'annotation fonctionnelle *in silico* de tels jeux de données requiert des approches fiables et de bonne qualité (Stenuit *et al.*, 2008).

Le dernier outil moléculaire, les biopuces ADN, semble de plus en plus adapté à l'étude de la structure des communautés microbiennes, mais également à la caractérisation de ses capacités métaboliques, et sera décrit plus précisément au sein de ce chapitre.

3. La technologie des biopuces ADN

3.1. Un outil de post-génomique

Apparues au milieu des années 90, après le séquençage complet des premiers génomes, les biopuces ADN ont été initialement mises au point pour l'étude simultanée de l'expression de tous les gènes d'un organisme (Schena *et al.*, 1995). Les biopuces ADN sont basées sur une hybridation des acides nucléiques dite inverse (comparée au Southern Blot) car elles utilisent des sondes immobilisées et ce sont les cibles qui sont marquées (Ehrenreich, 2006).

La première application des biopuce ADN en microbiologie de l'environnement date de 1997 (Guschin *et al.*, 1997). Cette première biopuce ADN était constituée de neuf sondes ciblant l'ADNr 16S, et permettait l'identification de bactéries dénitrifiantes. Depuis, l'utilisation des biopuces ADN s'est élargie, tant sur les domaines d'application que sur leur développement. Elles ont ainsi permis par exemple, dans le domaine de l'écologie microbienne, l'étude de la structure des communautés microbiennes d'écosystèmes variés, ou des potentialités métaboliques présentes au sein des environnements (He *et al.*, 2007; Rastogi *et al.*, 2010).

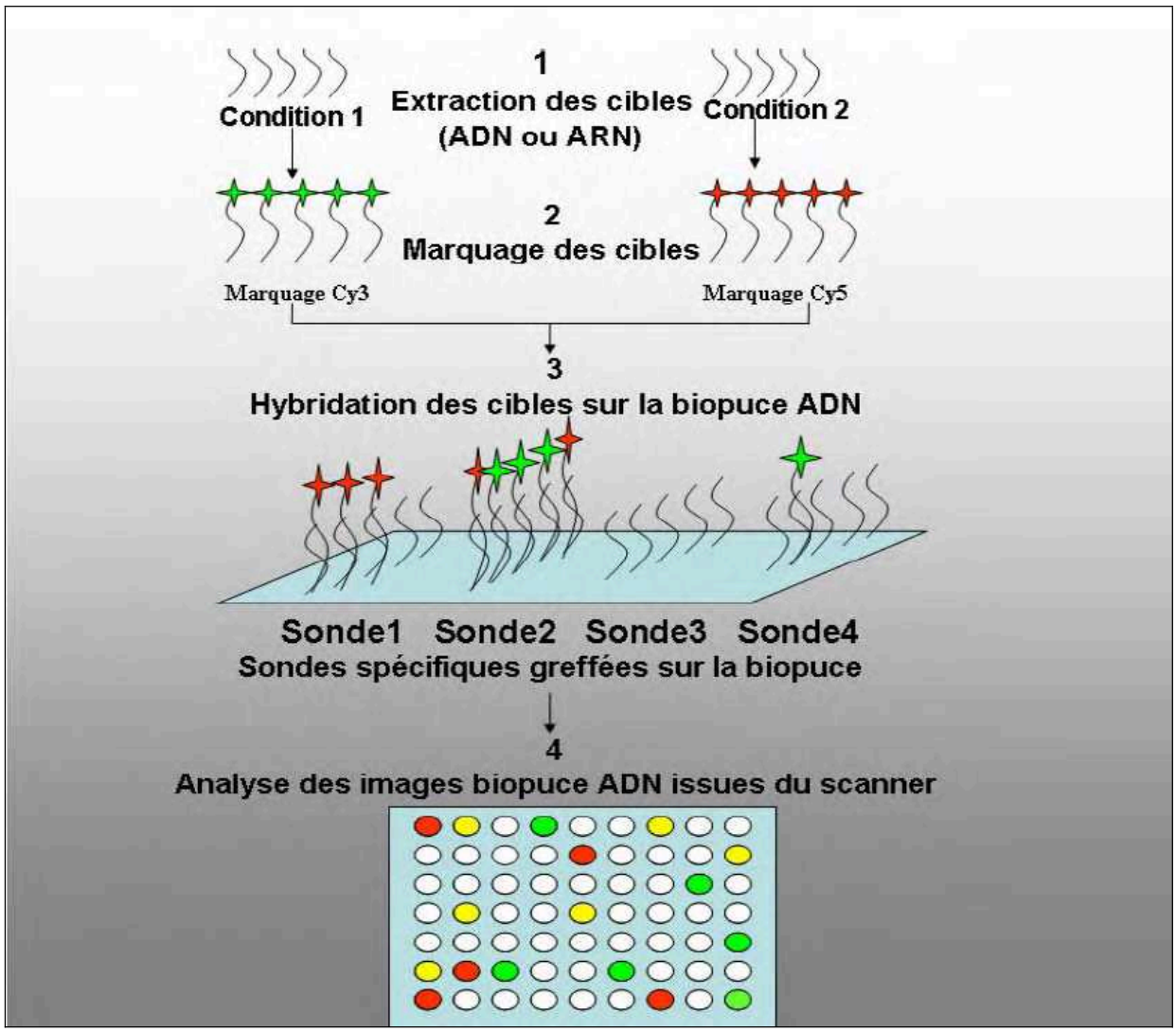


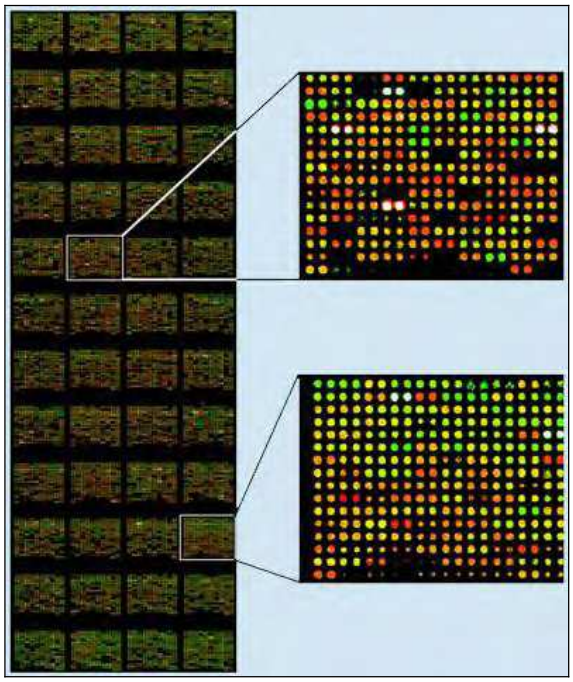
Figure 27 : Représentation schématique des différentes étapes d'une approche biopuce ADN. (Tirée de Joux *et al.*, 2010).

3.2. Principe

Comme décrit précédemment, le principe général des biopuces ADN repose sur une hybridation des acides nucléiques dite inverse (comparée au Southern Blot), car ce sont les sondes (ADN complémentaires, ADN génomique, produits PCR ou oligonucléotides) et non les cibles qui sont immobilisées sur un support solide (Figure 27) (Ehrenreich, 2006). La reconnaissance sonde-cible va se faire par complémentarité des bases constituant les acides nucléiques (Adénine et Tyrosine ou Uracile, Cytosine et Guanine) (Dufva, 2009b, a; Joux *et al.*, 2010). La fixation des sondes peut se faire sur un support solide, de type nitrocellulose (appelées macroarrays), ou de type lame de verre ou de silice (appelées microarrays) (Granjeaud *et al.*, 1999). Grâce aux nouvelles technologies de miniaturisation, il est possible de fixer sur une seule biopuce des centaines, voire même des dizaines de milliers de sondes différentes, augmentant ainsi la détection simultanée d'une grande variété de gènes ou de microorganismes.

Durant l'étape d'hybridation, plusieurs populations de cibles, marquées le plus souvent par des fluorochromes, peuvent être hybridées simultanément (Figure 27) (Dufva, 2009b, a). L'utilisation de plusieurs fluorochromes différents (généralement des cyanines Cy3 et Cy5) permet l'hybridation simultanée de plusieurs échantillons biologiques sur une même biopuce (Ehrenreich, 2006; Ying et Sarwal, 2009). La formation d'homoduplex (cible ADN), ou d'hétéroduplex (cible ARN) a lieu lorsqu'une sonde présente une complémentarité de séquence avec une cible (Galvão *et al.*, 2005). Les hybridations peuvent être faites de deux manières distinctes. La méthode développée initialement était basée sur une hybridation statique. Elle consistait à déposer des cibles dénaturées sur la surface de la biopuce dans un volume réduit (de quelques µL), puis ces cibles étaient recouvertes d'une lamelle de verre. Cette méthode permettait de former une environnement homogène d'hybridation tout en limitant le volume pour favoriser le contact avec les sondes (Ehrenreich, 2006; Dufva, 2009b; Joux *et al.*, 2010). Ce montage était alors placé dans une chambre d'hybridation hermétique pour être incubée à une température fixe dans un bain marie. Les temps d'hybridations étaient relativement longs (de 12 heures à 66 heures), car les constantes de diffusion de l'ADN dans l'eau sont extrêmement faibles (une molécule d'ADN par diffusion passive se déplacera de 2 à 3 mm en 24 heures), et ne favorisent donc pas les rencontres sondes-cibles (Sartor *et al.*, 2004; Joux *et al.*, 2010).

Afin d'améliorer cette étape, des automates allant du four à hybridation (Agilent : http://www.genomics.agilent.com/CollectionOverview.aspx?PageType=Application&SubPageType=ApplicationOverview&PageID=287), jusqu'à la station d'hybridation et de lavage



<u>Figure 28 :</u> Image d'une biopuce ADN obtenue après hybridation des cibles marquées et détection de la fluorescence à l'aide d'un scanner.
Chaque spot représente une hybridation entre une sonde spécifique et sa cible. Les nuances de couleurs (du noir au blanc) reflètent les niveaux d'intensité de fluorescence en relation avec la proportion de cibles capturées.

<u>Figure 29 :</u> Appareil HybLIVE^{TM} développé par l'entreprise Genewave.
Cet appareil permet un suivi en temps réel d'une hybridation ainsi que la réalisation de courbes de fusion.
Source (GeneWave) : http://www.genewave.com/products.php?id=8&produit=HybLive

multi biopuces complètement automatisée (Ventana : http://www.ventanadiscovery.com/) ont été développés. Ces systèmes automatisés assurent une agitation efficace du mélange, favorisant ainsi la mise en contact des sondes et des cibles. Il s'agit par exemple de l'advection chaotique (Hertzsch *et al.*, 2007), ou des systèmes de rotation de la chambre durant l'hybridation (McQuain *et al.*, 2004). Enfin, quelle que soit la méthode appliquée pour effectuer l'étape d'hybridation, il est ensuite indispensable de passer par des étapes de lavage pour éliminer les cibles non appariées ou mal appariées (hybridations aspécifiques) (Ehrenreich, 2006).

Les duplex sonde/cible formés lors de l'étape d'hybridation sont ensuite détectés, le plus souvent à l'aide d'un scanner pour biopuce ADN. Un faisceau laser va balayer toute la surface de la biopuce, excitant les fluorochromes de la cible, qui vont émettre de la lumière (Ying et Sarwal, 2009; Joux *et al.*, 2010). Cette lumière sera alors collectée par un photomultiplicateur permettant la transformation du signal lumineux en signal électrique, pour recréer une image numérique représentant la biopuce. Les images sont sauvegardées au format TIFF (Tagged Image File Format), sans perte d'information ou de qualité (les pixels ayant une gamme d'intensité dynamique allant de 0 à 65 535) (Figure 28). Concernant l'acquisition des données, une innovation originale a été proposée par la société Genewave (Marcy *et al.*, 2008). Le système HybLIVE™ (Figure 29) permet ainsi de réaliser une acquisition en temps réel de l'hybridation et d'effectuer des courbes de fusion. Une telle approche facilite la différenciation des hybridations parfaites de celles présentant un ou plusieurs mésappariements. La fenêtre de lecture reste limitée à une zone carrée de 13,5mm de côté, permettant l'analyse d'environ 10 000 sondes.

Les images des biopuces ADN, quel que soit le système utilisé, sont alors analysées à l'aide d'outils informatiques dédiés, assurant la récupération des informations pour chaque sonde. La première étape consiste en la détection et la localisation de chaque sonde sur la biopuce. Le positionnement d'une grille, permettant d'individualiser chaque spot, peut se faire de manière manuelle, semi-automatique ou automatique selon le logiciel utilisé, et selon les fichiers générés pendant la fabrication des biopuces. Puis, les algorithmes informatiques vont extraire l'intensité des pixels composant le spot (généralement la moyenne des signaux des pixels, ou encore appelée la densité de fluorescence) (Dufva, 2009b). Il est cependant indispensable d'éliminer le bruit de fond technique (local ou global), correspondant le plus souvent aux signaux mesurés autour du spot. Enfin, une normalisation est indispensable pour des comparaisons inter ou intra lame (du fait des relations non linéaires entre les fluorochromes par exemple) (Dharmadi et Gonzalez, 2004; Dufva, 2009b; Joux *et al.*, 2010).

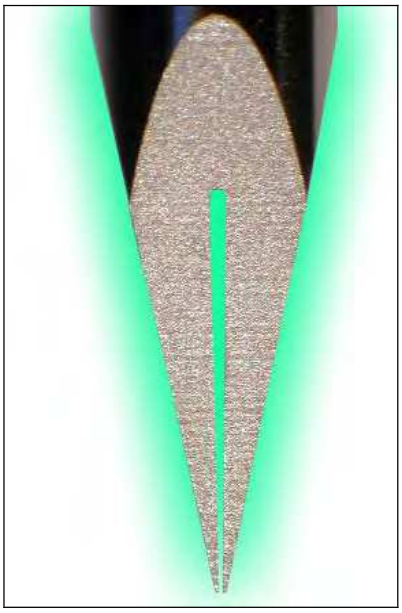


Figure 30 : Image microscopique d'une aiguille creuse utilisée classiquement pour réaliser des dépôts pour la fabrication de biopuces ADN *ex situ*.
Source (ArrayIT®): http://www.arrayit.com/.

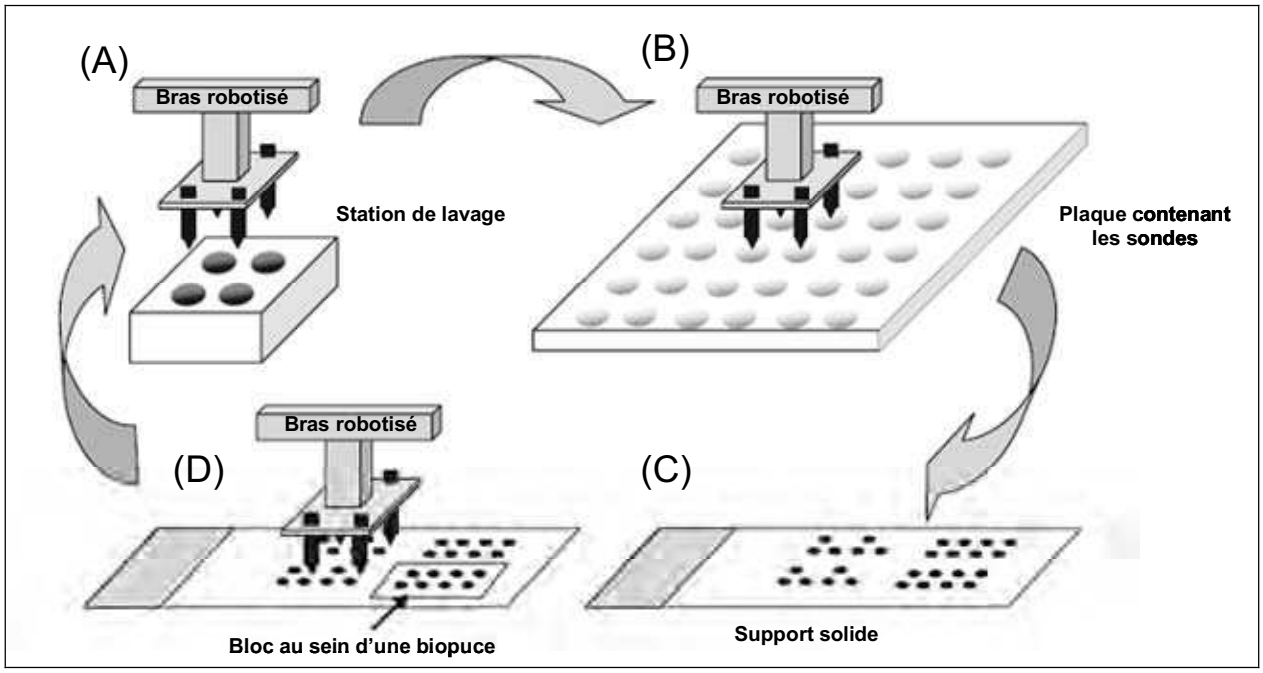


Figure 31 : Principe de fabrication de biopuces dites *ex situ*.
Les aiguilles de dépôt fixées à un bras robotisé vont tout d'abord être plongées dans une solution de lavage (A), puis dans les solutions contenant les sondes à déposer (B). Le bras robotisé va ensuite se diriger vers le support solide (C) (de type lame de verre par exemple) où vont être déposées les sondes (par contact ou par projection) une ou plusieurs fois (D). Le bras robotisé va ensuite se diriger vers la station de lavage pour rincer les aiguilles de dépôt, et commencer un nouveau cycle (tirée de Dufva, 2009a).

Les données seront alors interprétées en fonction de la question biologique posée (Ying et Sarwal, 2009).

3.3. Les différents types de sondes pour biopuces ADN

Deux types de sondes peuvent être utilisées pour les biopuces : les sondes ADN double brin dénaturées (produits PCR, ADNc et ADN génomiques), et les sondes oligonucléotidiques (Joux *et al.*, 2010). Les sondes ADNc et PCR ont été largement utilisées pour réaliser des biopuces transcriptomiques permettant le suivi de l'expression de gènes (Schena *et al.*, 1995). Ces sondes ont une très bonne sensibilité (de par leur taille moyenne de 500 pb) et, pour certaines études en conditions hétérologues, leur relative tolérance envers les variabilités polymorphiques est un avantage (Kreil *et al.*, 2006). Cependant, cette propriété peut s'avérer dans certains cas un inconvénient majeur, notamment pour l'étude de gènes d'une même famille et présentant de forte similarité de séquences. Enfin, les coûts engendrés par la préparation et la vérification de ces sondes sont importants. Les sondes ADN génomiques, quant à elles, sont généralement utilisées pour des biopuces permettant la comparaison de génomes proches ou pour le suivi de certains génomes dans des environnements complexes. Une alternative à ces sondes ADN double brin est l'utilisation des sondes oligonucléotidiques (Kreil *et al.*, 2006). Ces sondes, faciles à synthétiser, ont une spécificité relativement grande, en raison de leur taille réduite (20 à 70-mers), mais leur sensibilité est plus faible que les sondes ADN double brin (Kreil *et al.*, 2006; Joux *et al.*, 2010). De plus, et contrairement aux sondes ADN double brin, il est indispensable de connaître la séquence du gène ciblé pour pouvoir créer une sonde adaptée.

En fonction de la nature des sondes, les biopuces peuvent être construites de deux façons. La première consiste en une synthèse indépendante des sondes, suivie de leur fixation sur le support solide à l'aide d'un robot spotter (ce sont les biopuces dites *ex situ*). Dans ce cas, les aiguilles de dépôt sont plongées dans la solution de sondes, et ces dernières sont déposées par contact ou par projection sur la surface de la biopuce (Figures 30 et 31) (Joux *et al.*, 2010). Cette méthode peut être utilisée pour tout type de sondes, et les liaisons (covalentes ou non) entre la surface et les sondes varient (Dufva, 2009a). Par exemple, pour les sondes oligonucléotidiques, de nombreuses chimies de surface des lames sont disponibles, mais les surfaces les plus souvent utilisées sont recouvertes de groupements aminosilane (interaction électrostatique), époxy ou aldéhyde (fixations covalentes des sondes). Afin de réaliser la liaison covalente entre la lame et la sonde pré synthétisée, cette dernière possède souvent une modification sur l'une de ses extrémités. Le plus souvent il s'agit d'un groupement amine

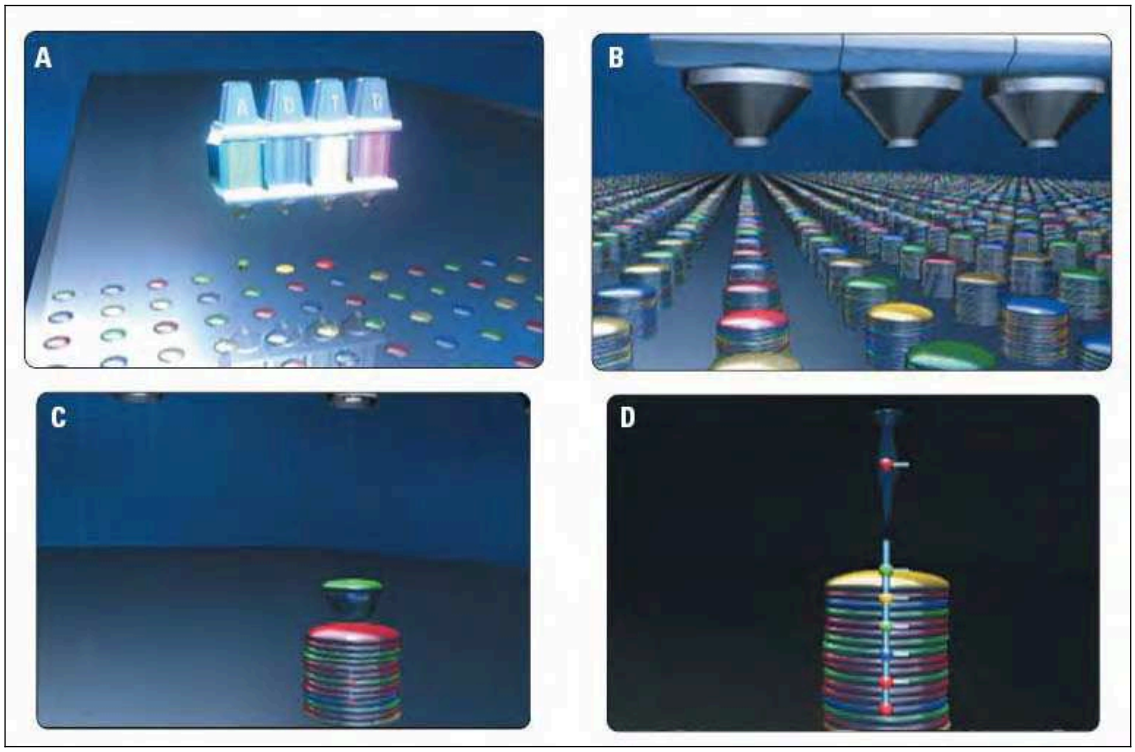


Figure 32 : Représentation schématique de la technologie de fabrication *in situ* des biopuces ADN Agilent.
(**A**) : La première couche de nucléotides est déposée sur le support solide. (**B**) : La synthèse continue de chaque oligonucléotide en fonction de sa séquence par projection. (**C**). Ajout d'une nouvelle base à un oligonucléotide. (**D**) Oligonucléotide totalement synthétisé.
Source (Agilent) : http://www.home.agilent.com/agilent/home.jspx?cc=US&lc=eng.

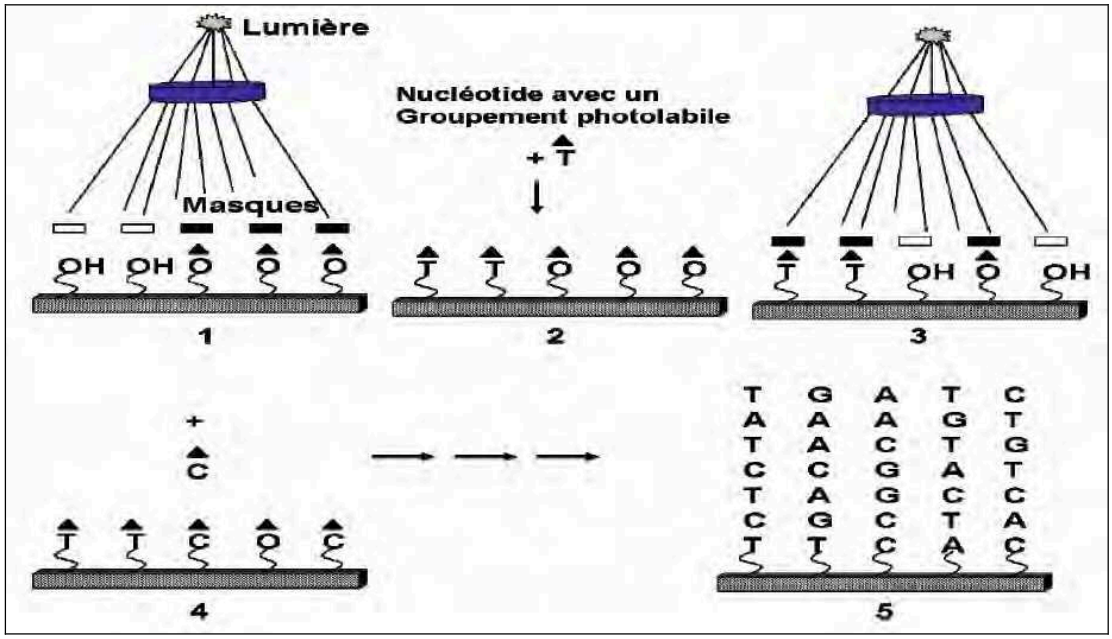


Figure 33 : Schéma de synthèse *in situ* des oligonucléotides par un procédé de photolithographie utilisé par Affymetrix.
Plusieurs jeux de masques permettent l'activation des groupements photolabiles (triangle noir). Lorsque que le jeu de masque laisse passer la lumière, le groupement photolabile de la chimie de surface ou des nucléotides est activé et permet la formation de la liaison covalente pour l'enchaînement des nucléotides (tirée de Joux *et al.*, 2010).

situé en 5' qui réagit avec les groupements époxy ou aldéhyde. La seconde façon de construire une biopuce ADN consiste à réaliser directement la synthèse des sondes sur la lame. Cette synthèse *in situ*, uniquement utilisable pour les sondes oligonucléotidiques, peut être réalisée par plusieurs méthodes. Dans le premier cas, les nucléotides sont projetés contre la lame selon une technologie analogue à celle d'une imprimante jet d'encre (Figure 32) (technologie Agilent). La seconde méthode de synthèse *in situ*, utilisée par l'entreprise Affymetrix®, permet d'obtenir des biopuces oligonucléotidiques très denses contenant plusieurs centaines de milliers de sondes et repose sur la photolithographie (Figure 33).

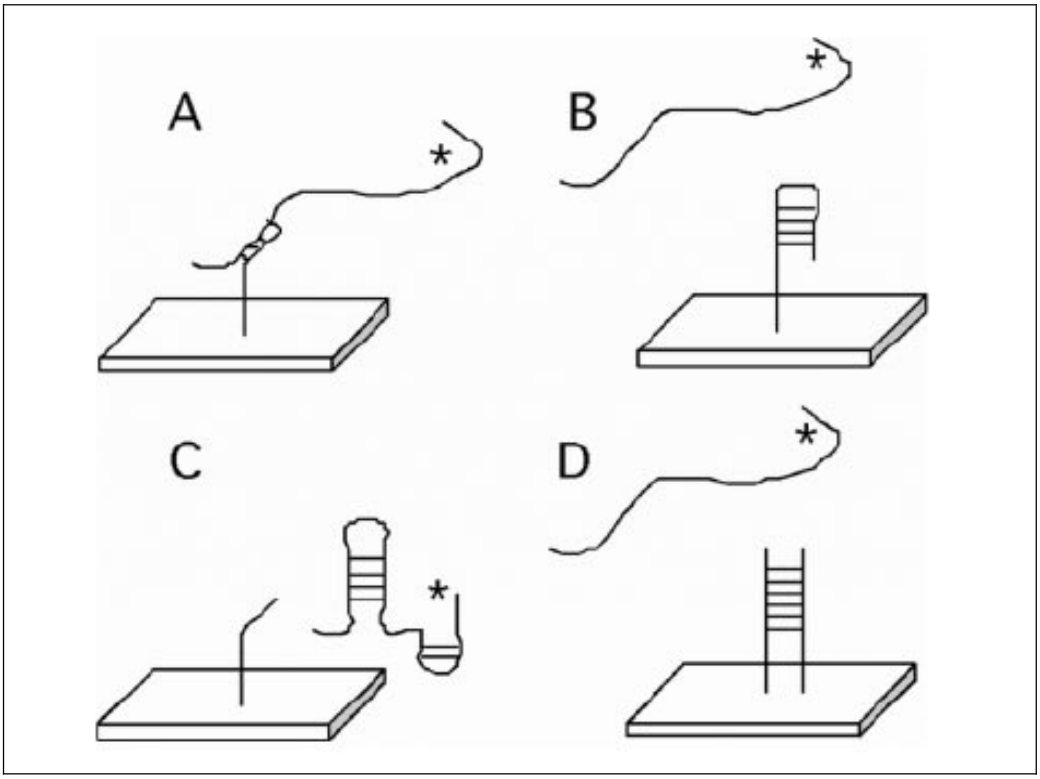
Avec l'augmentation de la qualité des oligonucléotides synthétisés, leur faible coût et leur excellente spécificité, les biopuces ADN composées d'oligonucléotides sont les plus utilisées actuellement (Kreil *et al.*, 2006; Joux *et al.*, 2010). Le point crucial dans la conception de biopuces à oligonucléotides sera donc la détermination des sondes comme nous le verrons dans le paragraphe suivant.

4. La détermination des sondes pour biopuces ADN

L'utilisation des biopuces ADN dans le domaine de l'écologie microbienne a pour but d'étudier et de mieux appréhender les communautés microbiennes présentes au sein d'un écosystème d'intérêt. L'efficacité de la biopuce ADN repose sur l'efficacité des sondes sélectionnées. Celles-ci doivent être sensibles, et reconnaître spécifiquement les groupes ciblés, même ceux faiblement présents dans l'échantillon étudié (Gentry *et al.*, 2006; Dufva, 2009b).

4.1. Spécificité et sensibilité des sondes oligonucléotidiques

D'après les critères de Kane (Kane *et al.*, 2000), un oligonucléotide est considéré comme spécifique (c'est-à-dire sans phénomènes d'hybridations croisées) s'il ne présente pas plus de 75 % d'identité avec l'ensemble des séquences ne correspondant pas à la séquence ciblée (condition 1), et s'il ne partage pas plus de 15 nucléotides consécutifs avec ces mêmes séquences (condition 2). L'utilisation d'oligonucléotides de taille importante, bien que diminuant le risque de valider la condition 1, augmente la probabilité de vérifier la condition 2. Inversement, l'utilisation de sondes courtes, bien que permettant d'éliminer les risques de présenter plus de 15 nucléotides consécutifs communs avec les cibles de l'échantillon, augmente le risque de valider la condition 1. Ceci n'est vrai que jusqu'à un certain seuil de taille, où la sonde devient trop courte pour être spécifique (Rimour *et al.*, 2005). Cependant, les critères définis par Kane semblent trop stringents pour des sondes d'une gamme de taille



<u>Figure 34 :</u> Schéma de différentes structures secondaires au niveau des sondes ou des cibles pouvant influencer l'efficacité d'appariement.

La cible marquée (visible avec un *) peut former un duplex avec la sonde (A). Il est également possible qu'une structure dite tige-boucle se forme au niveau de la sonde (B). Des structures secondaires peuvent également se former au niveau de la cible (C). Enfin, des dimères de sondes peuvent se former au niveau des spots (D). Ces trois cas (A, B et C) peuvent diminuer l'efficacité d'appariement (tirée de Pozhitkov *et al.*, 2006).

entre 20 et 25-mers et sont donc plus appropriés pour des sondes oligonucléotidiques longues (Kane *et al.*, 2000; Relogio *et al.*, 2002; He *et al.*, 2005; Rimour *et al.*, 2005).

De même, le type de mésappariements et la position du mésappariement au sein du duplex sonde/cible n'auront pas la même influence sur l'efficacité de l'hybridation. En effet, un mésappariement en position centrale déstabilisera fortement la formation de l'appariement. Ces propriétés sont utilisées par l'entreprise Affymetrix (http://www.affymetrix.com/estore/) pour la détermination de sondes permettant une discrimination des hybridations croisées. Cette société propose ainsi des biopuces ADN composées de sondes 25-mers. Un groupe de sonde est déterminé pour une cible avec une sonde montrant une hybridation parfaite avec la cible (sonde PM pour perfect Match) et des sondes formant un mésappariement en position centrale (sondes MM pour Mismatch). Il est ainsi facile de distinguer les hybridations parfaites des hybridations croisées. Cependant l'influence des mésappariement apparaît significative sur des sondes courtes mais devient plus faible avec l'augmentation de la taille de la sonde (Deng *et al.*, 2008). De plus, les forces d'appariement des bases étant différentes (G-C > A-T > G•G > G•T ≥ G•A > T•T ≥ A•A > T•C ≥ A•C ≥ C•C), certains mésappariement seront donc plus difficiles à discriminer. Ainsi la base G est capable de s'apparier fortement avec toutes les bases (SantaLucia et Hicks, 2004). Il est donc judicieux de privilégier des mésappariement mettant en jeu l'adénine. Plusieurs calculs thermodynamiques (comme l'énergie libre de liaison sonde/cible) permettent d'estimer la formation et la stabilité des duplex sondes/cibles en présence de mésappariements (Gresham *et al.*, 2010).

Afin de faciliter les interprétations avec une notion quantitative, il est également préférable de sélectionner des sondes montrant une même efficacité d'hybridation des cibles. Pour cela, la majorité des logiciels de détermination de sondes sélectionne des oligonucléotides partageant une température de fusion équivalente (Tm) (Lemoine *et al.*, 2009). De nombreuses méthodes de calcul existent pour déterminer ce Tm, la plus fréquente étant celle du plus proche voisin (ou Nearest-Neighbor), avec les paramètres de SantaLucia ou de Rychlik (Rychlik *et al.*, 1990; SantaLucia, 1998). Un paramètre étroitement lié à ce Tm est le pourcentage de guanine et de cytosine au sein des oligonucléotides (ou appelé %GC). (He *et al.*, 2005; Lemoine *et al.*, 2009). Néanmoins, ces calculs ne restent précis que pour des sondes en solution, et non pour des sondes greffées sur un support solide. De même, la formation de structures secondaires, soit au niveau des sondes soit au niveau des cibles, vont influencer les efficacités d'appariement (Figure 34). En effet, de telles structures vont empêcher ou limiter la formation des duplex sondes/cibles, réduisant ainsi la sensibilité des sondes (Matveeva *et al.*, 2003; Pozhitkov *et al.*, 2006). Il est cependant possible, grâce à des

calculs thermodynamiques, d'évaluer la stabilité de chaque structure secondaire en calculant son énergie libre de Gibbs. Ces calculs peuvent permettre de choisir les sondes montrant des énergies d'appariement suffisamment fortes pour déstabiliser ces structures secondaires (Pozhitkov *et al.*, 2007). Cependant, il est difficile d'utiliser des programmes de prédiction de structures secondaires comme MFOLD (Zuker, 2003), ou OligoWalk (Lu et Mathews, 2008), car les temps de calcul sont très importants. Par exemple, dans le cas d'un gène de 730 pb, le calcul des énergies pour chaque sous séquence de 19 bases a nécessité près d'une heure (Lu et Mathews, 2008).

Dans le cas de l'utilisation de cibles provenant de transcription inverse, la position ciblée par la sonde est également un paramètre important à prendre en compte. (Kreil *et al.*, 2006; Lemoine *et al.*, 2009). En effet, en raison des arrêts prématurés de la transcriptase inverse (mauvaise efficacité, structures secondaires), il est généralement conseillé de déterminer des sondes au plus proche de la région où est initiée la réaction de transcription inverse.

Lors de la détermination des sondes, il est également primordial d'exclure les régions de basse complexité pouvant être retrouvées au sein d'un grand nombre de séquences non ciblées (Wernersson et Nielsen, 2005; Kreil *et al.*, 2006). Plusieurs systèmes permettent de filtrer ces régions, et sont généralement intégrés aux logiciels de définition de sondes (Lemoine *et al.*, 2009). Par exemple, l'outil BLAST (souvent intégré aux logiciels de design de sondes), possède le programme DUST pour s'affranchir de ces régions de basse complexité.

Toutefois, tous ces critères pris séparément ne sont pas suffisants pour définir les sondes oligonucléotidiques les plus spécifiques et les plus sensibles. C'est pourquoi, il serait indispensable de tous les prendre en compte pour définir les meilleures sondes possibles (He *et al.*, 2005; Liebich *et al.*, 2006; Deng *et al.*, 2008; Gresham *et al.*, 2010). Cependant, la thermodynamique des hybridations des acides nucléiques sur support solide est loin d'être élucidée, ce qui rend difficile la détermination des meilleures sondes oligonucléotidiques pour biopuces ADN (Li *et al.*, 2005; Pozhitkov *et al.*, 2007).

4.2. Systèmes d'implémentation pour la sélection de sondes

De très nombreux logiciels et programmes pour déterminer des sondes pour biopuces ADN ont été développés, qu'ils soient commerciaux, ou libres d'utilisation. Dans ce paragraphe, nous ne nous intéresserons qu'aux logiciels libres d'accès et d'utilisation, et dont les caractéristiques sont connues, définies, et accessibles par tous les utilisateurs.

<u>Tableau 10 :</u> Logiciels de détermination de sondes oligonucléotidiques pour biopuces ADN et critère de recherche pour optimiser leur qualité. (Adapté de Lemoine *et al.*, 2009).

Logiciel	Référence	Recherche d'hybridations croisées	Régions de basse complexité	Position de la sonde sur la cible	Nombre de sondes déterminées par gène
ArrayOligoSelector	(Bozdech *et al.*, 2003)	BLAST Calculs thermodynamiques	Elimination des sondes via un calcul de score	Classement des sondes selon la distance par rapport au 3'	Choix de l'utilisateur
CommOligo	(Li *et al.*, 2005)	Calculs thermodynamiques et critères de Kane	Masque de séquences	Design débuté soit en 3' ou 5' (utilisateur)	Choix de l'utilisateur
HPD	(Chung *et al.*, 2005)	Alignement multiple et clustering hiérarchique	Non considéré	Non considéré	Toutes les sondes correspondant aux critères
OliD	(Talla *et al.*, 2003)	BLAST	Elimination des sondes via l'étude des répétitions	Préférence des sondes en position 3'	Choix de l'utilisateur
OligoArray 2.0	(Rouillard *et al.*, 2003)	BLAST Calculs thermodynamiques	Masque de séquences (nucléotides répétés)	Distance par rapport au 3' définie (utilisateur, max 1500)	Choix de l'utilisateur
OligoFaktory	(Schretter et Milinkovitch, 2006)	BLAST	Masque de séquences par DUST	Design débuté soit en 3' ou 5' (utilisateur)	Choix de l'utilisateur (maximum de 3 sondes)
OligoPicker	(Wang et Seed, 2003)	BLAST et recherche de séquences répétées	Masque de séquences par DUST	Design uniquement en position 3' ou 5' (utilisateur)	Choix de l'utilisateur (maximum de 5 sondes)
OligoWiz	(Wernersson et Nielsen, 2005)	BLAST Calculs thermodynamiques	Elimination des sondes via un calcul de score	Score de localisation selon 3', 5' ou centre (utilisateur)	Toutes les sondes correspondant aux critères
Oliz	(Chen et Sharp, 2002)	BLAST Critères de Kane	Non considéré	Détermination uniquement en position 3'	Toutes les sondes correspondant aux critères
Osprey	(Gordon et Sensen, 2004)	Système propre basé sur les matrices de score	Masque de séquences (nucléotides répétés)	Biais 5' ou 3' en dernier paramètre de sélection	Toutes les sondes correspondant aux critères
PICKY	(Chou *et al.*, 2004)	Méthode 'Suffix Array' avec critères de Kane et calculs thermodynamiques	Méthode 'Suffix Array' pour les régions de basse complexité	Non considéré	Choix de l'utilisateur (maximum de 5 sondes)
PRIMEGENS	(Xu *et al.*, 2002)	BLAST, critères de Kane Alignement multiple	Non considéré	Non considéré	Choix de l'utilisateur
PROBEmer	(Emrich *et al.*, 2003)	Méthode 'Suffix Array'	Non considéré	Intervalle de position optionnel	Toutes les sondes correspondant aux critères
Probesel	(Kaderali et Schliep, 2002)	Méthode 'Suffix Array' Calculs thermodynamiques	Non considéré	Non considéré	Une sonde par gène
ProbeSelect	(Li et Stormo, 2001)	Méthode 'Suffix Array' Calculs thermodynamiques	Elimination des sondes (répétitions de nucléotides)	Non considéré	Choix de l'utilisateur
ROSO	(Reymond *et al.*, 2004)	BLAST	Elimination des sondes (répétitions de nucléotides)	Intervalle de position optionnel	Choix de l'utilisateur
TherMODO	(Leparc *et al.*, 2009)	BLAST	Masque de séquences (nucléotides répétés)	Non considéré	Toutes les sondes correspondant aux critères
YODA	(Nordberg, 2005)	Système propre (SeqMatch)	Masque de séquences (séquences interdites)	Choix de l'utilisateur	Toutes les sondes correspondant aux critères

4.2.1. Systèmes d'implémentation et critères de spécificité des sondes

Sur la totalité des programmes existants, une grande majorité utilise l'outil informatique BLASTn pour déterminer les hybridations croisées potentielles, comme ArrayOligoSelector (Bozdech *et al.*, 2003), OliD (Talla *et al.*, 2003), OligoPicker (Wang et Seed, 2003), OligoWiz (Wernersson et Nielsen, 2005), ROSO (Reymond *et al.*, 2004) ou OligoArray 2.0 (Tableau 10) (Rouillard *et al.*, 2003; Lemoine *et al.*, 2009). A partir des résultats produits par l'analyse BLASTn, ces logiciels utilisent généralement les critères de Kane décrits plus haut pour confirmer ou non l'existence d'une réaction croisée potentielle (Kane *et al.*, 2000). C'est notamment le cas des outils : ArrayOligoSelector, OligoArray, OligoPicker, OligoWiz et Oliz (Tableau 10). D'autres outils tentent d'éviter les problèmes potentiellement rencontrés avec l'approche BLASTn. En effet, en abaissant au maximum la valeur du mot permettant d'initier l'alignement local entre deux séquences (W=7), une telle approche peut empêcher la détermination d'hybridations croisées réelles. Ainsi, en raison de l'algorithme, deux séquences identiques à l'exception d'une base tous les 7 nucléotides ne pourront donc pas être alignées bien qu'elles présentent entre elles un pourcentage élevé de similitude. De plus, un alignement local ne reflète pas réellement l'identité moyenne entre deux séquences. C'est pourquoi plusieurs solutions ont été mises en place, comme l'utilisation d'un alignement global (pour CommOligo), ou une nouvelle approche nommé 'Suffix Array' (pour PICKY, PROBEmer, Probesel et ProbeSelect). Le principe original de Suffix Array consiste en la création d'une base recensant, par ordre alphabétique, de courtes séquences appelées suffixes, ainsi que leur position au sein du jeu de séquences fourni par l'utilisateur (Chou *et al.*, 2004). Cette base sert en fait de référence pour réaliser toutes les recherches de sondes. Il est donc possible, en s'appuyant sur ces sous séquences déjà référencées, de réaliser une recherche rapide et précise. Certains outils se basent, en plus, sur des calculs thermodynamiques (calculs des énergies libres pour le duplex sonde/cible, pour la formation de structures secondaires, etc.…) pour déterminer les duplex sondes/cibles les plus stables, et donc les sondes sensées être les plus sensibles (comme ArrayOligoSelector, OligoArray et OligoWiz ou Probesel, ProbeSelect et PICKY).

Deux autres approches originales sont : celle utilisée par HPD, qui se base sur un alignement multiple et un clustering hiérarchique, et celle appliquée par YODA, qui utilise un système de recherche proche de BLAST, nommé SeqMatch, pour définir des sondes spécifiques (Chung *et al.*, 2005; Nordberg, 2005). SeqMatch se base sur l'utilisation de deux algorithmes différents, l'un après l'autre. Cette approche est conçue pour identifier rapidement si la sonde considérée peut engendrer des hybridations croisées au sein du jeu de

séquences fourni. Tout d'abord, la recherche a lieu avec une taille de mot à 4, garantissant de retrouver toutes les hybridations croisées avec un seuil d'identité supérieur à 80 %. Si cette recherche ne donne rien, une seconde recherche exhaustive a lieu avec un seuil d'identité inférieur à 80 %, prenant plus de temps pour identifier les hybridations croisées. SeqMatch a l'avantage de s'arrêter dès qu'une hybridation croisée est détectée, pour passer à la sonde suivante. Au final, SeqMatch donne une liste des sondes ne présentant aucune hybridation croisée au sein du jeu de séquences fourni. Il est possible que YODA ne retrouve aucune sonde, nécessitant alors une nouvelle recherche en modifiant les paramètres.

En ce qui concerne l'élimination des régions de basse complexité, un grand nombre de programmes utilisent des filtres ou des masques pour les exclure. La prise en compte, ou non, de telles régions peut cependant être laissée à l'utilisateur pour certains outils, comme YODA. Enfin, il est également possible de détecter ces zones soit par des calculs particuliers sur le jeu de séquences fourni (prenant en compte le nombre d'occurrences d'un pattern simple au sein de ces séquences) (comme ArrayOligoSelector ou OligoWiz), soit par les propriétés définies par la méthode 'Suffix Array' (pour PICKY). Certains se basent aussi sur l'outil BLAST pour filtrer ces régions de basse complexité en utilisant par exemple le programme DUST pour s'affranchir de ces régions.

La position des régions ciblées sur le gène est également prise en compte dans plusieurs programmes de design. De nombreux outils favorisent la position 3' des gènes, comme Oliz. En effet, cet outil est spécialement dédié à l'étude des régions 3' UTR (région non traduite des ARN). D'autres laissent l'utilisateur choisir la zone à considérer (5', centrale ou 3') (comme OligoPicker ou ROSO), ou fournissent tous les résultats et laissent le choix des sondes (comme YODA).

Enfin, le nombre de sondes défini, par gène, peut aussi être considéré comme un critère de spécificité. Certains logiciels donnent toutes les sondes répondant aux critères considérés, comme HPD, YODA ou OligoWiz (Tableau 10), Probesel, quant à lui, ne détermine qu'une sonde par gène (Kaderali et Schliep, 2002), et d'autres laissent à l'utilisateur la possibilité de définir le nombre de sondes par gène à déterminer (comme OligoPicker, OligoFactory et PICKY). Dans ce cas, certains outils sélectionnent les sondes approchant au mieux les paramètres définis s'il y a trop de sondes (comme PICKY), d'autres s'arrêtent lorsque le nombre de sondes est atteint. Enfin, cet objectif peut également ne pas être atteint si les paramètres sont trop stringents.

Tableau 11 : Autres critères et approches utilisés par les logiciels pour la détermination de sondes oligonucléotidiques. (Adapté de Lemoine *et al.*, 2009).

Logiciel	Longueur de la sonde	Détermination du Tm	%GC	Calculs de structures secondaires
ArrayOligoSelector	Fixée par l'utilisateur	Non considéré	Elimination des sondes selon le seuil fixé par l'utilisateur	Alignement de la sonde avec sa séquence inverse complémentaire
CommOligo	Fixée par l'utilisateur (de 10 à 128-mers)	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Filtration des séquences selon le critère fixé par l'utilisateur	Alignement de la sonde avec sa séquence inverse complémentaire
HPD	Fixée par l'utilisateur	Non considéré	Elimination des sondes selon le seuil fixé par l'utilisateur	Calculs thermodynamiques (formation de tige/boucle)
OliD	Fixée par l'utilisateur	Non considéré	Elimination des sondes selon le seuil fixé par l'utilisateur	Basé sur MFOLD
OligoArray 2.0	Fixée par l'utilisateur (de 15 à 75-mers)	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Elimination des sondes selon le seuil fixé par l'utilisateur	Basé sur un module spécifique comparable à MFOLD
OligoFaktory	Fixée par l'utilisateur	Intervalle choisi par l'utilisateur	Non considéré	Non considéré
OligoPicker	Fixée par l'utilisateur (de 20 à 100-mers)	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Non considéré	Alignement de la sonde avec sa séquence inverse complémentaire
OligoWiz	Fixée par l'utilisateur (basée sur le Tm)	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Non considéré	Evalué par un algorithme spécifique au programme
Oliz	Fixée à 50-mers	Intervalle fixe vers 76°C, calcul par l'outil 'prima' du package EMBOSS	Intervalle fixe compris entre 45 % et 50 %	Non considéré
Osprey	Fixée par l'utilisateur (de 10 à 90-mers)	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Non considéré	Calculs des énergies libres d'appariement et de tige/boucle
PICKY	Fixée par l'utilisateur	Optimisation du Tm, calculs selon la méthode du plus proche voisin	Intervalle modifié sur le critère du Tm	Inclus dans la vérification des hybridations croisées
PRIMEGENS	Fixée par l'utilisateur	Intervalle choisi par l'utilisateur, basé sur Primer3	Intervalle choisi par l'utilisateur, basé sur Primer3	Calculs basés sur données de Primer3
PROBEmer	Fixée par l'utilisateur	Intervalle choisi par l'utilisateur	Elimination des sondes selon le seuil fixé par l'utilisateur	Alignement de la sonde avec sa séquence inverse complémentaire
Probesel	Fixée par l'utilisateur	Calcul par la méthode du plus proche voisin	Non considéré	Basé sur MFOLD
ProbeSelect	Fixée par l'utilisateur	Non considéré, calcul du Tm selon une méthode spécifique au programme	Non considéré	Alignement de la sonde avec sa séquence inverse complémentaire
ROSO	Fixée par l'utilisateur	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Intervalle conseillé entre 40 % et 65 %	Calculs d'énergie libre d'homoduplex et de tige/boucle
TherMODO	Fixée par l'utilisateur	Non considéré	Non considéré	Nombreux calculs thermodynamiques spécifiques
YODA	Fixée par l'utilisateur	Intervalle choisi par l'utilisateur, calcul par la méthode du plus proche voisin	Elimination des sondes selon le seuil fixé par l'utilisateur	Alignement de la sonde avec sa séquence inverse complémentaire

4.2.2. Systèmes d'implémentation et critères de sensibilité des sondes

La taille des sondes influence leur sensibilité, c'est pourquoi la plupart des logiciels décrits dans le Tableau 11 permettent soit de définir une même taille pour chaque sonde, ou tout du moins une gamme de taille restreinte. Seul l'outil Oliz impose une taille unique de 50 nucléotides (Chen et Sharp, 2002).

Le calcul du Tm, quant à lui, est réalisé soit directement par le logiciel (comme OligoPicker ou ProbeSelect), soit via d'autres sous-programmes intégrés, comme prima, du package EMBOSS, implémenté dans Oliz (Rice *et al.*, 2000). Dans les deux cas, les Tm sont définis par rapport à une gamme, qu'elle soit imposée par le logiciel, ou modulable par l'utilisateur. Si le Tm est considéré comme un critère prioritaire à respecter, d'autres paramètres seront alors automatiquement ajustés par les logiciels comme la taille des sondes (c'est le cas pour OligoArray, OligoPicker, YODA et PICKY). Certains outils n'effectuent pas de calcul de Tm, bien que ce critère soit pris en compte indirectement par l'évaluation du %GC de chaque sonde (pour ProbeSelect ou ArrayOligoSelector par exemple). Enfin, les logiciels OligoFaktory, OligoPicker, OligoWiz, Osprey, Probesel et ProbeSelect font complètement abstraction de ce paramètre (Tableau 11).

La caractérisation de structures secondaires potentielles est également réalisée par plusieurs logiciels. Les premiers, comme ArrayOligoSelector, CommOligo, OligoPicker, PROBEmer, ProbeSelect et YODA, se basent sur la recherche de courtes sous-séquences au sein de la séquence de la sonde, pouvant engendrer la formation de structures tige-boucle (c'est-à-dire complémentaires l'une avec l'autre) (Lemoine *et al.*, 2009). Les logiciels comme OliD, OligoArray et Probesel quant à eux, se basent sur des calculs thermodynamiques réalisés avec le programme MFOLD, pour déterminer l'énergie libre de Gibbs d'une ou plusieurs structures (comme par exemple le duplex sonde/cible, les structures tige-boucle au sein de la cible et/ou de la sonde, la formation d'homo dimères pour la sonde, etc.…). Enfin, pour limiter les temps de calculs, HPD, OligoWiz et ROSO utilisent des systèmes propres pour définir ces paramètres thermodynamiques. Enfin, PICKY couple les calculs thermodynamiques (détermination d'homo dimères ou de structures tige-boucle) à la recherche d'hybridations croisées (Chou *et al.*, 2004).

4.2.3. Systèmes d'implémentation et adaptabilité

Ces différents outils sont accessibles de différentes manières. Ainsi, certains sont disponibles en téléchargement directe, et utilisables en local (comme ArrayOligoSelector, HPD, OligoArray, OligoFaktory, OligoPicker, Oliz, PICKY, ProbeSel, ProbeSelect, ROSO,

YODA et PRIMEGENS) sur divers systèmes d'exploitation (Windows, Linux ou MacOS). De plus, certains sont uniquement utilisables à travers une interface Web, comme PROBEmer ou Osprey. Les plus adaptables sont OligoArray, OligoWiz, PICKY, PRIMEGENS et YODA qui peuvent s'utiliser en local sur ces trois systèmes d'exploitation. Il est important de noter que certains ne sont disponibles qu'après demande auprès des auteurs, ce qui peut rendre difficile la récupération de l'outil (comme CommOligo, OliD, PICKY, ProbeSel ou ProbeSelect), comme Lemoine le précise pour OliD (Lemoine *et al.*, 2009). Toujours si l'on se base sur cette étude récente, ces outils montrent des temps de calcul très variables pour un même design de sondes (le test consiste en la définition d'une sonde pour chacun des 1 421 gènes testés, connus pour être impliqués dans le développement du système nerveux de la souris). Ainsi, CommOligo nécessite 1 156 minutes, et ROSO 418 minutes pour ce design, alors que YODA ne nécessite que 3 minutes, mettant en évidence une grande variabilité des performances des outils testés.

De plus, les outils de détermination des sondes doivent être le plus généralistes possibles, pour permettre de tester la spécificité de chaque sonde, en fonction de la composition de l'échantillon étudié. Ainsi, par souci de flexibilité la plupart des programmes permettent néanmoins à l'utilisateur de s'appuyer sur des bases de données personnelles dont les séquences sont fournies au logiciel sous la forme d'un fichier FASTA (ArrayOligoSelector, OligoArray, OligoPicker, OSPREY, PICKY, PROBEmer, ROSO et YODA). D'autres outils, quant à eux, permettent l'élaboration d'une base de référence à façon en utilisant les données de RefSeq (comme c'est le cas de Mprime) ou de GenBank (pour OligoFaktory). Seul OligoWiz ne permet pas d'utiliser, pour la recherche des hybridations croisées potentielles, d'autres bases que celles disponibles sur le site. Sur demande, il est cependant possible de pouvoir ajouter les séquences des génomes de nouveaux organismes (Wernersson et Nielsen, 2005). A l'heure actuelle, il n'existe cependant pas d'outils s'appuyant sur l'ensemble des données disponibles dans les bases de données généralistes (génomes complets, métagénomes, etc.…) (Pozhitkov *et al.*, 2007; Lemoine *et al.*, 2009).

5. Les différents types de biopuces ADN et leurs applications en écologie microbienne

Les biopuces ADN peuvent être appliqués à de nombreux domaines d'étude des microorganismes (Ehrenreich, 2006; Gentry *et al.*, 2006; Joux *et al.*, 2010). On peut ainsi les classer en quatre grands types de biopuces : les biopuces génomiques (ou CGA : Community

Genome Arrays), les biopuces transcriptomiques, les biopuces phylogénétiques (ou POA : Phylogenetic Oligonucleotide Arrays) et les biopuces fonctionnelles (ou FGA : Functional Gene Arrays). Ces dernières feront l'objet d'une attention plus particulière, et une description des applications de ce type de biopuces en bioremédiation sera réalisée.

5.1. Les biopuces génomiques

Du fait des transferts horizontaux de gènes entre les souches bactériennes, les microorganismes, bien que montrant une forte parenté phylogénétique (séquence d'ADNr 16S proches), peuvent cependant faire apparaître d'importantes différences au niveau de l'organisation et de la composition génique de leurs génomes (Joux *et al.*, 2010). La comparaison de leurs génomes respectifs pourrait donc permettre de comprendre l'origine génétique de nombreuses différences phénotypiques, mais aussi d'étudier les mécanismes responsables de l'évolution des génomes microbiens.

Plusieurs types de biopuces dites génomiques existent. Par exemple, les biopuces appelées WGA (Whole Genome Array) sont composées de sondes générées par amplification PCR, de chaque gène du génome considéré et peuvent présenter une densité allant jusqu'à 50 000 produits PCR sur une lame de verre. Il est cependant nécessaire, pour le développement des biopuces WGA, de sélectionner des couples d'amorces pour chaque ORF, et donc d'amplifier chaque gène. De plus, du fait de la taille de ces produits PCR, des problèmes d'hybridations non spécifiques peuvent apparaître. Une alternative de fabrication est l'utilisation de sondes oligonucléotidiques spécifiques de chaque ORF, dont la spécificité doit cependant être vérifiée. Récemment, l'utilisation de ce type de biopuce se tourne vers ce que l'on appelle les « tilling arrays » (Gresham *et al.*, 2008). Il est ainsi possible d'obtenir une couverture complète d'un génome avec de petits oligonucléotides permettant théoriquement de détecter les réarrangements, les insertions, les inversions, les duplications et les polymorphismes d'une simple base. Il est également possible de fabriquer une biopuce génomique, même si la séquence génomique des souches d'intérêt n'est pas connue (Cho et Tiedje, 2001). Dans cette expérience, quatre souches de *Pseudomonas* ont permis, après fragmentation de leurs génomes, clonage et amplification PCR, d'obtenir de 60 à 96 sondes par souche. Cette biopuce a démontré ses capacités d'identification et de discrimination de parenté entre souches bactériennes.

Un nouveau format de biopuce génomique appelé CGA (Community Genome Array) permet de s'affranchir d'une construction laborieuse et coûteuse des biopuces décrites précédemment. Pour les CGA, les génomes entiers de bactéries cultivables sont fixés

directement sur une lame de verre. Pour exemple, c'est le cas d'une étude ciblant 67 génomes bactériens (Wu *et al.*, 2004). L'ADN de ces microorganismes a été isolé et utilisé comme sondes. En utilisant des températures d'hybridation élevées (55 à 75°C), et l'ajout de formamide (jusqu'à 50 % du volume d'hybridation), les auteurs ont réalisé des différenciations au niveau de la souche (par exemple entre la souche d'*Azoarcus tolulyticus* Td-21, et les autres souches testées de cette même espèce) et validé cette biopuce. Cette biopuce a également servi pour comparer plusieurs écosystèmes différents (quatre sédiments marins, trois sédiments de rivière et trois sols). Ils ont pu différencier ces trois environnements par l'obtention des signatures spécifiques à chacun de ces écosystèmes. Ces résultats ont de plus démontré le potentiel de ce type de biopuce, d'un point de vue spécificité et sensibilité, pour caractériser les microorganismes présents dans divers écosystèmes (Wu *et al.*, 2004; Gentry *et al.*, 2006)

Ces biopuces ADN sont très discriminantes au niveau du genre et de l'espèce. De plus, elles peuvent être utilisées pour caractériser rapidement les parentés qui existent entre les espèces connues et celles nouvellement isolées (Gentry *et al.*, 2006). Cependant, seuls les génomes des souches connues et cultivées peuvent être utilisés pour générer de telles biopuces. Néanmoins, une alternative est l'utilisation de sondes déduites d'études de métagénomique. Ainsi, récemment, à partir d'informations obtenues par métagénomique d'environnements marins, des sondes oligonucléotidiques de 70-mers ont été déterminées (Rich *et al.*, 2008). Ce type de biopuces a permis ainsi d'identifier et de suivre à la fois la dynamique, et l'activité microbienne d'environnements complexes (dans le cas de l'étude réalisée, d'un écosystème marin).

5.2. Les biopuces transcriptomiques

Ces biopuces ADN ont été les premières a être utilisées pour évaluer l'expression de tous les gènes d'un organisme (Schena *et al.*, 1995). Leur spécificité consiste à déposer sur la biopuce des sondes spécifiques de chacun des gènes d'un organisme donné. Ce type de sondes peut être conçu avec des produits PCR, ou des oligonucléotides. Cela permet de comparer la variation de l'expression de la totalité des gènes d'un microorganisme dans diverses conditions d'une manière très simple.

En écologie microbienne, ces approches peuvent également être menées. Récemment, l'analyse de la réponse aux stress oxydants d'un *consortia* a été réalisée (Scholten *et al.*, 2007). Ce *consortia*, composé de quatre espèces anaérobies (deux bactéries syntrophiques : *Desulfovibrio vulgaris* et *Synthrophobacter fumaroxidans*, et deux archées : *Methonosarcina*

barkeri et *Methanospirillum hungatei*) a été étudié après un stress oxydatif. L'analyse des données obtenues avec ou sans stress oxydatif a pu montrer l'expression différentielle de nombreux gènes (70, 107, 86 et 96, respectivement), et ce pour chacune des souches de manière précise. Ces travaux ont démontré l'efficacité de ce type d'outil pour l'étude de l'expression des gènes d'un *consortia*, malgré la complexité de l'échantillon, ce qui en fait potentiellement un outil efficace pour l'étude des communautés au sein de leur environnement naturel.

5.3. Les biopuces phylogénétiques

Les biopuces phylogénétiques (ou POA) sont actuellement le type de biopuces ADN le plus largement utilisé pour étudier les communautés bactériennes (Joux *et al.*, 2010). Grâce aux sondes oligonucléotidiques, ciblant les gènes codant l'ADNr 16S, les POA permettent notamment la discrimination des groupes procaryotiques. En effet, le biomarqueur phylogénétique qu'est l'ARNr 16S permet de discriminer aisément les différents microorganismes, et est généralement utilisé pour l'analyse de la structure des communautés bactériennes retrouvées dans divers écosystèmes. Les zones des ARNr 16S, ciblées par les sondes, sont choisies en fonction du degré de résolution attendu. Les séquences hautement conservées sont utilisées pour une détermination à des niveaux taxonomiques supérieurs (famille, ordre, classe) alors que les séquences hypervariables peuvent différencier les espèces et les genres.

De nombreuses POA ont été développées ciblant un ou plusieurs groupes bactériens, comme la puce « SRP-PhyloChip » (ciblant les communautés sulfato-réductrices procaryotiques) (Loy *et al.*, 2002), ou la puce « RHC-PhyloChip », ciblant les bétaprotéobactéries de l'ordre des *Rhodocyclales* (Loy *et al.*, 2005), ou encore celle développée pour cibler les alphaprotéobactéries afin de suivre la rhizosphère du maïs (Sanguin *et al.*, 2006). Cependant, certaines biopuces récentes et plus complètes permettent d'avoir une vision plus globale des communautés microbiennes. La PhyloChip (Brodie *et al.*, 2006; DeSantis *et al.*, 2007) est ainsi composée de 297 851 sondes ciblant 8 741 taxons (bactéries et archées). Cette dernière a été développée pour suivre l'évolution des communautés microbiennes d'un point de vue spatial au sein des sols antarctiques (Yergeau *et al.*, 2008), afin d'étudier les effets du réchauffement climatique sur les communautés. Une autre étude, en utilisant cette même biopuce, a permis de démontrer une très grande diversité microbienne au sein de mines d'uranium (Rastogi *et al.*, 2010).

Cependant, cette biopuce ne permet de cibler que les séquences d'ADNr 16S connues (Brodie *et al.*, 2006; DeSantis *et al.*, 2007; Joux *et al.*, 2010). Une nouvelle avancée majeure dans l'étude des communautés microbiennes est la création de sondes exploratoires pour l'étude de l'ensemble des communautés microbiennes. C'est avec cet objectif, de pouvoir identifier des séquences non encore répertoriées dans les bases de données, qu'une première biopuce prototype exploratoire a été développée (Militon *et al.*, 2007). De telles sondes, permettant de cibler des séquences potentiellement affiliées à un genre donné, sont définies à l'aide du logiciel nommé PhylArray. Cette nouvelle génération de POA permettra donc d'appréhender la biodiversité microbienne dans sa globalité, et sans a priori sur les séquences du gène ciblé (Joux *et al.*, 2010)..

5.4. Les biopuces fonctionnelles

Les biopuces fonctionnelles (ou FOA) sont le plus souvent appliquées à l'étude de gènes impliqués dans les grands cycles biogéochimiques ou dans divers procédés de bioremédiation (Gentry *et al.*, 2006; Joux *et al.*, 2010). Les sondes greffées sur ces biopuces sont soit des produits PCR (entre 200 et 1000 nucléotides environ), soit des oligonucléotides (de 15 à 70-mers) (Sessitsch *et al.*, 2006). Une des premières études mises en place avec des FOA pour suivre l'expression de gènes au sein d'un environnement en présence de xénobiotiques a été réalisée en 2003 (Dennis *et al.*, 2003). La biopuce utilisée cible un total de 64 gènes, dont certains sont connus pour être impliqués dans la dégradation de composés aromatiques chlorés (comme l'acide 2,4-dichlorophénoxyacétique), ou d'autres composés (comme le naphtalène ou le carbazole), ou encore des gènes dits « constitutifs » (comme l'ARNr 16S). Les sondes greffées ont été obtenues par amplification PCR des gènes d'intérêt (de 271 à 1 238 nucléotides). Cette FOA leur a permis d'étudier l'expression des gènes ciblés sur des extraits de rejets de boues d'une usine de papeterie, amendés ou non, par une résine acide. Cette étude a mis en évidence l'expression de plusieurs gènes impliqués dans la dégradation de composés aromatiques chlorés. Cependant, la limite principale de cette FOA est qu'elle ne permet pas d'évaluer la totalité de la variabilité génique des gènes ciblés, et nécessite donc d'être enrichie pour étudier des environnements complexes (Stenuit *et al.*, 2008).

L'augmentation de la qualité des oligonucléotides synthétisés, et leur très bonne spécificité en fait un outil idéal pour appréhender la grande variabilité des gènes fonctionnels au sein d'un écosystème, et donc différencier les cibles d'une manière plus précise qu'avec des produits PCR. L'utilisation de FOA développées en utilisant des oligonucléotides s'est

donc accrue ces dernières années. Un exemple est le développement d'une biopuce constituée de sondes oligonucléotidiques de 50-mers, ciblant 2 402 gènes connus, dont les produits sont respectivement impliqués dans la biodégradation de molécules aromatiques, ou dans la résistance aux métaux (Rhee *et al.*, 2004). De plus, ces sondes,selon les régions ciblées des gènes, peuvent s'hybrider à plusieurs variants d'un même gène, ou être spécifiques d'un seul gène (Rhee *et al.*, 2004). La plupart des sondes choisies ont d'abord été validées d'un point de vue spécificité et sensibilité, via des tests avec des souches pures. Les oligonucléotides de 50-mers définis permettent ainsi de différencier les cibles possédant en général moins de 88 % de similarité avec leurs sondes, si les conditions d'hybridation sont suffisamment stringentes (T = 50° C ; %Formamide = 50%). De plus, et même si leur sensibilité est inférieure à celle des produits PCR, elles permettent tout de même de détecter leurs cibles à partir de 5 à 10 ng d'ADN génomique pur ou dans 50 à 100 ng d'ADN génomique en mélange. Les auteurs ont ainsi démontré la faisabilité et l'efficacité d'une telle biopuce oligonucléotidique pour l'étude de microcosmes et de sols enrichis.

Une approche originale de l'utilisation des biopuces ADN a également été réalisée pour l'étude de la diversité de la monooxygénase impliquée dans la dégradation du benzène (Iwai *et al.*, 2007). Pour cette étude, les auteurs ont développé une première biopuce oligonucléotidique de 87 sondes (60-mers), ne ciblant que la diversité des gènes codant cette enzyme, en se basant à la fois sur les séquences présentes au sein des bases de données, et sur les séquences qu'ils ont isolé des environnements étudiés. Ils ont montré que la biopuce développée ne permettait d'appréhender qu'une faible partie de la diversité des monoxygénases du benzène, et ont donc fait évoluer leur biopuce en isolant de nouvelles séquences de plusieurs écosystèmes, augmentant le nombre de sondes total à 150 (Iwai *et al.*, 2008; Iwai *et al.*, 2010).

Cependant, jusqu'à récemment, les FOA ne présentaient pas une forte densité, de par la difficulté de la détermination des sondes. Le développement d'une biopuce haute densité appelée GeoChip, ciblant plus de 10 000 gènes impliqués dans les cycles du carbone, de l'azote, du soufre, du phosphore mais aussi dans la dégradation de divers polluants et de la résistance aux métaux a depuis été réalisée (He *et al.*, 2007). De nombreuses études ont été réalisées avec cette FOA, dont certaines sur des environnements contaminés par des polluants aromatiques (Liang *et al.*, 2009a; Liang *et al.*, 2009b). La première étude, réalisée sur un sol contaminé par du pétrole, a permis de montrer que certains gènes impliqués dans la dégradation de polluants (comme le biphényle) ou de métabolites clés (comme le catéchol ou le protocatéchuate) sont fortement représentés à un niveau moyen de pollution. La seconde

étude porte sur l'ozonation *ex situ* d'un sol pollué par divers hydrocarbures (monoaromatiques, polycycliques et aliphatiques), puis d'une étape de biodégradation avec et sans bioaugmentation (Liang *et al.*, 2009b). L'utilisation de la GeoChip, dans ce cas, montre une diminution du nombre de gènes détectés, après ozonation. Cependant, avec l'étape de biodégradation (avec ou sans bioaugmentation), le nombre de gènes total mesuré augmente de nouveau, bien qu'un changement de la communauté microbienne ait eu lieu, d'après les résultats d'hybridation pour chaque gène. Cette étude montre donc également l'intérêt de l'utilisation d'une telle FOA pour suivre les impacts des traitements de bioremédiation d'environnements pollués sur les communautés microbiennes.

Une limite majeure de la majorité de ces études est l'utilisation de l'ADN et non de l'ARN, pour réaliser les analyses ou les suivis. Cela est lié à la difficulté d'obtention du matériel génétique en qualité et en quantité suffisante, ce qui est d'autant plus vrai pour les cibles ARN dont l'extraction reste souvent problématique à partir d'environnements complexes (Gentry *et al.*, 2006). Des hybridations avec ce type de molécules ont cependant déjà été entreprises. L'une d'entre elles, a notamment permis de démontrer la présence et l'expression de gènes impliqués dans la dégradation du naphtalène au sein de microcosmes de sols enrichis (Rhee *et al.*, 2004). Une autre étude plus récente, utilisant une biopuce fonctionnelle composée de 2 006 oligonucléotides de 50-mers ciblant de nombreux gènes, dont ceux impliqués dans la résistance et la réduction des métaux lourds, a permis d'étudier l'expression de ces gènes au sein d'un environnement complexe pollué (Wu *et al.*, 2006; Gao *et al.*, 2007). En effet, après amplification linéaire des ARN environnementaux cibles (extraits d'eaux contaminées par des hydrocarbures, de l'uranium, des solvants organiques et du nitrate), ceux-ci ont été hybridés sur la biopuce (Gao *et al.*, 2007). Cette approche a permis d'éliminer les biais liés à l'utilisation de l'ADN en étudiant uniquement les communautés actives de l'écosystème d'intérêt. Malheureusement, à l'heure actuelle la majorité des études réalisées se font avec des cibles de nature ADN, permettant uniquement de déterminer la présence ou non des gènes étudiés, sans information sur leur niveau d'expression et donc sur leur régulation potentielle au cours d'un processus biologique d'intérêt.

6. Conclusion

L'augmentation des applications des biopuces ADN dans les études d'écologie microbienne démontre leur fort potentialité, et laisse supposer qu'elles continueront d'être très informatives pour appréhender le fonctionnement des écosystèmes, malgré l'apparition des

approches de séquençage à très haut débit (Ehrenreich, 2006; Stenuit *et al.*, 2008). Ceci est d'autant plus vrai que cette technologie est en constante évolution, permettant notamment d'avoir des biopuces avec une densité de plus en plus importante (deux millions de sondes, et prochainement quatre millions annoncées par l'entreprise Nimblegen), et des formats adaptés pour analyser un grand nombre d'échantillons biologiques (comme celui proposé par la société Nimblegen, avec une biopuce composée de 12 blocs de 135 000 sondes chacun, chaque bloc pouvant être hybridé indépendamment). De plus, l'accumulation des données de séquence permet d'améliorer la détermination de sondes spécifiques et sensibles. Cependant, de nombreux écueils doivent être franchis, comme l'évaluation rapide de l'efficacité des sondes déterminées (Kreil *et al.*, 2006; Joux *et al.*, 2010). En outre, même avec les bases de données en constante évolution, des sondes exploratoires doivent être définies pour identifier l'entière diversité des communautés microbiennes et de leurs capacités métaboliques dans les différents écosystèmes.

Conclusion Générale

Les sols sont les premiers écosystèmes touchés par des pollutions. Une majorité de ces contaminations est liée à la présence d'hydrocarbures, principalement des HAP. Ces polluants touchent également les communautés procaryotiques présentes au sein de ces environnements, dont la diversité est très importante. Cependant, ces communautés sont capables de dégrader ces composés récalcitrants. C'est pourquoi l'utilisation de leurs capacités à dépolluer les sols se développe de plus en plus, en remplacement de techniques plus onéreuses et plus invasives. Toutefois, les mécanismes métaboliques mis en jeu dans ces processus de dépollution restent largement méconnus. Les seules données disponibles décrivent les voies de dégradation de HAP modèles comme le naphtalène ou le phénanthrène, en grande majorité pour des microorganismes isolés.

Il est donc indispensable de mieux appréhender les potentialités métaboliques des microorganismes épurateurs, généralement organisés en *consortia*, afin d'améliorer les techniques de bioremédiation. Dans ce cadre, l'utilisation d'outils de post-génomique comme les biopuces ADN, pour analyser ces écosystèmes dans leur globalité, semble très approprié. En effet, cette technique peut permettre de caractériser les voies métaboliques, leurs régulations et les relations métaboliques qui peuvent exister entre les différents organismes. Dans ce cadre, l'utilisation d'outils de post-génomique comme les biopuces ADN, pour analyser ces écosystèmes dans leur globalité semble très approprié. Cependant, une des limitations actuelles de cette approche est la détermination des sondes, qui ne ciblent que les gènes dont les séquences ont été caractérisées. La fraction inconnue des microorganismes, qui est largement majoritaire, reste donc ignorée.

Des travaux de recherche ont donc été réalisés afin de développer une biopuce fonctionnelle exploratoire, pour évaluer et déterminer les capacités métaboliques microbiennes d'un écosystème de type sol contaminé par des HAP. Les résultats de ces travaux seront présentés au cours de ce mémoire dans la partie Résultats.

MATERIEL ET METHODES

<u>Tableau 12 :</u> Analyse des hydrocarbures présents au sein de l'écosystème sol étudié.
Ces données ont été fournies par l'entreprise BioBasic Environnement. Toutes les valeurs sont indiquées en mg/ kg de masse sèche. Les méthodes d'analyses utilisées sont confidentielles. Aucune donnée n'est disponible pour les hydrocarbures aliphatiques.

Hydrocarbures totaux (C_{10}-C_{40})	**5 400**
Hydrocarbures totaux (C_{10}-C_{12})	**800**
Hydrocarbures totaux (C_{12}-C_{16})	**1 200**
Hydrocarbures totaux (C_{16}-C_{21})	**2 000**
Hydrocarbures totaux (C_{21}-C_{35})	**1 400**
Hydrocarbures totaux (C_{35}-C_{40})	**72**
Hydrocarbures Aromatiques Polycycliques (HAP)	
Naphtalène	**620**
Acénaphthylène	**110**
Acénaphthène	**14**
Fluorène	**47**
Phénanthrène	**430**
Anthracène	**160**
Fluoranthène	**270**
Pyrène	**210**
Benzo(a)anthracène	**79**
Chrysène	**74**
Benzo(b)fluoranthène	**93**
Benzo(k)fluoranthène	**36**
Benzo(a)pyrène	**70**
Dibenzo(ah)anthracène	*<9*
Indeno(1,2,3-cd)pyrène	**37**
Benzo(ghi)perylène	**43**
Somme des 16 HAP	**2 300**
Hydrocarbures aromatiques monocycliques	
Benzène	*<1,1*
Toluène	**5,2**
Ethylbenzène	*<1,1*
m, p Xylène	**12**
o Xylène	**6,9**
Xylènes totaux	**18,9**
Cumène	*<1*
p, m Ethyltoluène	**2,7**
Mesitylène	**4,8**
o Ethyltoluène	*<1,1*
Pseudocumène	**11**
Somme	**43**

1. Matériel Biologique

1.1. *Sphingomonas paucimobilis* sp. EPA505

La souche bactérienne EPA505 provient du fournisseur DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH). Elle est fournie sous forme lyophilisée. Après réhydratation de la souche bactérienne dans 1ml de milieu riche LB (Luria Bertani) suivant le protocole donné par le fournisseur, du glycérol est ajouté à une concentration finale de 20 %, puis la suspension est préparée en aliquots de 100µL avant d'être congelée à -80°C jusqu'à utilisation.

1.2. Terre contaminée

L'écosystème étudié de type sol est pollué par des hydrocarbures variés (Tableau 12). Cette terre a été fournie par l'entreprise Biobasic Environnement et a été stockée dès réception à -150°C. Le protocole d'extraction et d'analyse des HAP utilisé est confidentiel, car il est propriété de l'entreprise Biobasic Environnement, de même que l'origine et les caractéristiques du sol. La présence de nombreux polluants aromatiques comme des BETX (représentant environ 432 mg/kg de sol sec), mais surtout des composés aromatiques de type HAP (environ 2 300 mg/kg de sol sec pour les seize HAP principaux) a pu être détectée. C'est pourquoi cet écosystème a été choisi : de par sa forte pollution en HAP, comme le naphtalène (620 mg/kg de sol sec), le phénanthrène (430 mg/kg de sol sec), le fluoranthène (270 mg/kg de sol sec), le pyrène (210 mg/kg de sol sec), l'anthracène et l'acénaphthylène (respectivement 160 et 110 mg/kg de sol sec).

2. Caractérisation de la souche *Sphingomonas paucimobilis* sp. EPA505

2.1. Précultures bactériennes en milieu liquide

Les précultures sont réalisées dans un volume de 70mL de milieu LB additionné de streptomycine à une concentration finale de 100µg/mL (Sigma-Aldrich). L'ensemencement est réalisé avec 100µL de suspension de la souche EPA505, puis incubées à 37°C pendant 24h, à l'abri de la lumière, sous une agitation de 160 rpm grâce à un agitateur Rotatest (Bioblock Scientific).

2.2. Cultures sur milieux minéraux en présence de différents polluants comme seule source de carbone

Trois polluants de type HAP (phénanthrène (Fluka), naphtalène (Fluka), fluoranthène (Aldrich)) ont été solubilisés dans de l'acétone (Merck) à une concentration finale de 2mg/mL. Chaque solution est préalablement filtrée sur des Minisart RC4, de pore 0,2µm (Sartorius). Dans des Erlenmeyers stériles de 250mL, un volume de 2mL de solution filtrée (soit un polluant seul, soit un mélange 50/50 de chaque) est tout d'abord déposé de manière stérile. Puis, les récipients sont laissés sous une hotte à flux vertical pendant une nuit pour évaporer l'acétone, et ainsi obtenir une répartition homogène du polluant. Un volume de 100mL de milieu minéral M457 (Annexe 1), additionné de Tween 80 (0,2g/L) est ensuite ajouté dans chaque Erlenmeyer. De la streptomycine est également ajoutée à une concentration finale de 100µg/mL.

Les précultures obtenues au bout de 24h d'incubation, décrites précédemment, sont alors centrifugées à température ambiante à 5000g, pendant 5min. Les surnageants sont éliminés et chaque culot est repris dans 1mL de milieu M457. Cette étape de centrifugation et de lavage est répétée une nouvelle fois pour éliminer les traces de milieu LB. Chaque erlenmeyer contenant 100mL de milieu M457 est enfin ensemencé avec 500µL de suspension bactérienne ainsi obtenue, sauf un qui servira de témoin négatif.

Ces cultures sont incubées à 28°C pendant 27h à l'abri de la lumière sous une agitation de 160 rpm. Des prélèvements sont réalisés à différents temps (0, 3, 6, 8, 10, 13, 17, 21, 24 et 27 h) pour mesurer par spectrophotométrie, à une longueur d'onde de 620nm, la croissance bactérienne, et à 405nm, l'activité métabolique (apparition de semi quinone). Une culture témoin est également réalisée dans les mêmes conditions, avec comme seule source de carbone et d'énergie du glucose à une concentration finale de 15g/L.

2.3. Suivi analytique de la biodégradation

2.3.1. Préparation des gammes étalons pour l'analyse CLHP

Les concentrations de chaque HAP étudié sont suivies durant le processus de biodégradation en cultures pures. Les gammes étalons sont réalisées à partir d'un mélange de naphtalène (0,01g/L), de phénanthrène (0,001g/L) et de fluoranthène (0,01g/L) dans 10mL d'acétonitrile 100 % (Sigma-Aldrich). Après complète dissolution, ce mélange est dilué au 1/100, au 1/250, au 1/500, au 1/750 et au 1/1000, pour obtenir les différents points de chacune des gammes. La préparation des gammes est réalisée trois fois, et de manière indépendante. Chaque mesure de dilution est réalisée en trois réplicats, et ce pour chaque gamme.

2.3.2. Préparation des échantillons

Un Erlenmeyer complet est utilisé pour chaque point du suivi réalisé. L'extraction des HAP présents est réalisée par l'ajout de 14mL de dichlorométhane (Sigma-Aldrich) dans la fraction liquide. Une émulsion entre la phase organique et la phase aqueuse est réalisée par agitation orbitale pendant 16h à 250rpm. Après une décantation de 30min et un léger dégazage, la phase aqueuse est éliminée et la phase organique est récupérée. Le dichlorométhane est alors évaporé sous une hotte à flux vertical. Enfin, chaque culot est alors repris dans 15mL d'acétonitrile 100 % (Sigma-Aldrich), puis le tout est filtré avec des Minisart RC4 de pore 0,2µm (Sartorius). L'extrait est conservé à -20°C.

2.3.3. Analyse CLHP

La quantification des polluants est réalisée par analyse CLHP, équipée d'une pompe (Waters® 600 controller), d'un passeur automatique (AS1000 Spectra Physics équipé d'une boucle d'injection de 20µL) et d'un détecteur (Waters® 2996 Photodiode Array Detector). Les mesures sont réalisées entre 210 et 400nm. Le débit d'analyse est de 1,5mL/min avec comme éluant un mélange acétonitrile 80 % (VWR), H_2O ultrapure 20 %. Les produits à analyser sont séparés sur une colonne de type C_{18} HAP (longueur de la colonne : 250mm, diamètre : 4,6mm, taille des particules : 5µm, référence commerciale : 186001265). Les données sont acquises et analysées à l'aide du programme Millenium®. Les temps de rétention ont été vérifiés à l'aide de solutions pures de HAP qui ont servi pour la réalisation des gammes étalons.

2.4. Extraction d'acides nucléiques

2.4.1. Extraction d'ADN total

Un aliquot de 10mL de préculture bactérienne incubée dans les conditions détaillées précédemment est centrifugé à 6000g pendant 5min à température ambiante. Le culot est lavé avec 1mL de tampon TE 1X (Tris Base 10mM, EDTA 1mM, pH 8.0) puis les bactéries sont de nouveau centrifugées à 2000g pendant 5min à température ambiante. Le culot bactérien est ensuite repris avec 500µL de TE 1X, puis plongé dans un bain-marie à ébullition pendant 15min et enfin centrifugé 15min à 4°C à 10 000g. Cette dernière étape est répétée trois fois. Le surnageant contenant l'ADN est récupéré, dosé par spectrophotomètre Nanodrop (Nanodrop) afin de déterminer la meilleure concentration à utiliser pour une bonne efficacité de PCR.

Tableau 13 : Amorces et description des conditions d'amplification utilisées pour la caractérisation des gènes d'intérêt.
Ces amorces ont été utilisées durant les étapes de caractérisation des gènes codant les enzymes clefs des voies de biodégradation des HAP, mais également pour générer les matrices ADN indispensables à la réalisation des gammes étalons de PCR quantitative. *: les amorces déterminées pour les gènes *xylX* et *nahD* ont été utilisées pour caractériser les séquences complètes des gènes *bphC* et *ahdA1c*. Nomenclature: **M**: A et C; **R**: A et G; **W**: A et T; **S**: G et C; **Y**: C et T; **K**: G et T; **V**: A, G et C; **H**: A, C et T; **D**: A, G et T; **B**: G, T et C; **I**: A, C, G et T.

Fragment recherché	Amorce Sens	Séquence (5' – 3')	Amorce Antisens	Séquence (5' – 3')	Température (°C) d'hybridation	Temps (sec) d'hybridation
Amorces utilisées pour amplifier les gènes cataboliques ciblés						
xylX – bphC*	X_R1	ACCTGCASCTTCCAGTTGCC	C_R1_d	CKYTCRTTRCARTGCATRAA	45	45
bphC	C_F1_d	GAYYTBTGGCAYCAYCGCAT	C_R1_d	CKYTCRTTRCARTGCATRAA	45	30
bphC – bphA3	C_F1_d	GAYYTBTGGCAYCAYCGCAT	A3_R3_d	TGRCAIGGRAAIGCYTT	40	30
bphA3	A3_F1_d	ACIGAYGGITAYCARGAY	A3_R3_d	TGRCAIGGRAAIGCYTT	40	30
bphA3 – ahdA2c	A3_F1_d	ACIGAYGGITAYCARGAY	A2c_R2_d	TSRTCIAYRTAYTTICC	40	30
ahdA2c	A2c_F1_d	GAYGAYGAYMGIYTIGAR	A2c_R3_d	ACCATCATIGTRTCDAT	40	30
ahdA2c – ahdA1c	A2c_F	CCTTATGACCCGCACTGACG	A1c_R3_d	GCYTCISYRTCYTCCAT	53	30
ahdA1c	A1c_F1_d	TGYGTITAYCAYCARTGG	A1c_R3_d	GCYTCISYRTCYTCCAT	40	30
*ahdA1c – nahD**	A1c_F1_d	TGYGTITAYCAYCARTGG	D_R2_d	TBCGIGCCTTGCGRTATTC	40	30
phnA1a – phnA2a	A1a_F1	CATCGCATTGCCATTAGTG	A2a_R1_d	CTTGACKAGYACCTCRCCRT	56	30
ahdA4	A4_F1_d	GWGCRAATCTKGCSGGTGG	A4_R1_d	ARCCMGCCTGCTTGAGSA	56	30
bphB	B_F1_d	TTYGGIAARYTBGAYGT	B_R1_d	GGIGCVACICCRTTVAC	50	40
nahE	E_F1	CCAAAACCGTCGATGTAGGT	E_R1_d	TATCGCGGGTGTTGATRCAG	50	45
Amorces utilisées pour générer les matrices PCR pour les étapes de RT-PCR quantitative						
bphC	C_F1	TGGGAGAGAAAGCAAATGG	C_R1	TAATGGAAGGCTCAACCGA	56	30
bphA3	A3_F1	GCTGACCTTCTACTGCGCCA	A3_R1	CGGTGTAGACGCAGTCCGAA	68	15
ahdA2c	A2c_F1	GGTTCCTTCGACATCGCCAC	A2c_R1	AACGGCGATCTTTGAGCGGA	68	15
ahdA1c	A1c_F1	GGACGCACCACAATCTACAAT	A1c_R1	TATCTTGCGGGTCATCGTG	60	30
phnA1a	A1a_F1	CATCGCATTGCCATTAGTG	A1a_R1	GGCGTCACCGGAACTTGTTC	56	30
phnA2a	A2a_F1	CAGAGCCGGTCCAAATATCG	A2a_R1	CTTGACTAGTACCTCGCCGT	59	30
ahdA4	A4_F1_d	GWGCRAATCTKGCSGGTGG	A4_R1_d	ARCCMGCCTGCTTGAGSA	56	30
bphB	B_F1_d	TTYGGIAARYTBGAYGT	B_R1_d	GGIGCVACICCRTTVAC	50	40

Gènes : *xylX* : sous-unité α de la toluène dioxygénase, *bphC* : dihydroxynaphtalène dioxygénase, *bphA3* : sous-unité de ferrédoxine, *ahdA2c* : petite sous-unité d'oxygénase, *ahdA1c* : grande sous-unité d'oxygénase, *nahD* : 2-hydroxychromène-2-carboxylate isomérase, *phnA1a* : sous-unité α de la dioxygénase initiale, *phnA2a* : sous-unité β de la dioxygénase initiale, *ahdA4* : ferrédoxine réductase, *bphB* : *cis*-dihydrodiol déshydrogénase et *nahE* : hydratase-aldolase

2.4.2. Extraction d'ARN totaux sur cultures en présence de différents polluants comme seule source de carbone

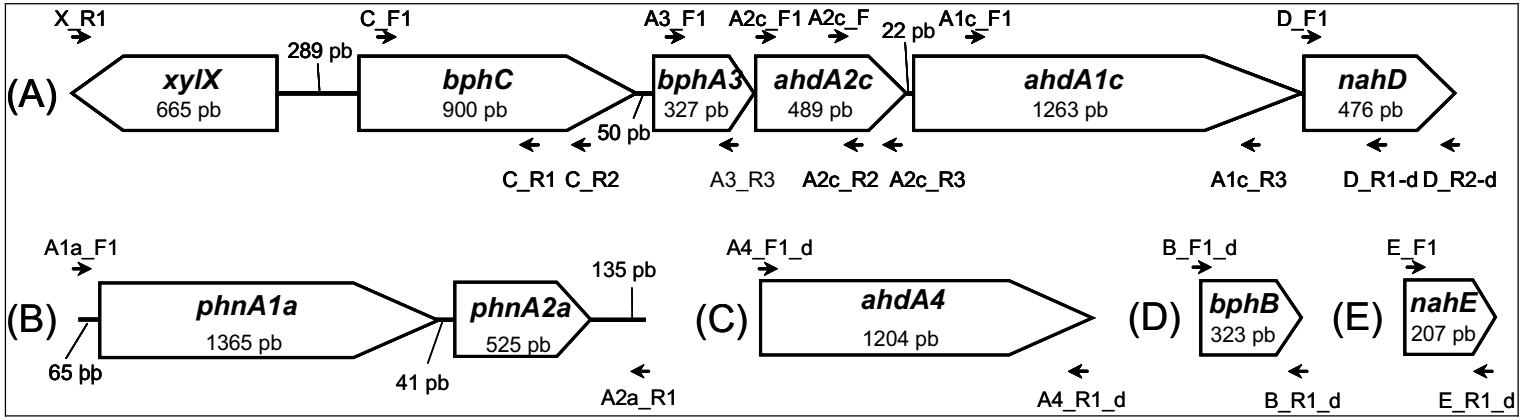
Les ARN totaux sont extraits à différents temps des cultures réalisées (0, 3, 6, 8, 10, 13, 17, 21, 24 et 27 h) grâce au kit RNeasy® Mini Kit (QIAGEN). Un aliquot de culture bactérienne de 10mL est immédiatement centrifugé à 6000g dans une Multifuge 3 SR (Heraeus) pendant 5min à 4°C. Les surnageants sont éliminés et les culots sont repris dans 600µL de tampon RLT (kit RNeasy®) additionné de β–mercaptoéthanol à 0,1 % final. La suspension bactérienne est introduite dans un tube eppendorf contenant 0,5g de billes de verre de 106µm de diamètre (Sigma-Aldrich) préalablement traitées et lavées. Deux broyages successifs de 30s sont réalisés à 30Hz, et à 4°C, à l'aide d'un broyeur à billes MM301 (Retsch). Les billes et les débris cellulaires sont ensuite sédimentés par centrifugation, et 350µL du surnageant sont prélevés, puis mélangés à 250µL d'éthanol absolu. La suspension est déposée sur une colonne du kit RNeasy®, puis centrifugé à 8000g à 4°C pendant 15s. Les étapes suivantes sont effectuées selon les recommandations du fournisseur (« Bacterial Protocol » du kit RNeasy®).

Les ARN totaux extraits sont alors traités avec 1,5 unités (U) de DNase I (Invitrogen) selon les recommandations du fournisseur et l'ajout de 20 unités (U) de RNasin Inhibiteur + (Promega). La qualité des ARN extraits est ensuite estimée à l'aide du kit RNA 6000 Nano (Agilent Technologies), et à l'appareillage associé (Bioanalyser 2100). L'estimation de la qualité est réalisée à partir de 2µL de solution d'ARN et selon les recommandations du fournisseur (Agilent). La concentration en acides nucléiques extraite est déterminée via un dosage spectrophotométrique (Nanodrop).

2.5. Caractérisation des gènes codant les enzymes clés des voies de biodégradation des HAP

2.5.1. Amplification PCR, clonage, séquençage des gènes ciblés

Toutes les réactions d'amplification sont effectuées dans un volume final de 50µL contenant : 20ng d'ADN total extrait de la souche EPA505, 0,5 unités (U) de GoTaq DNA polymérase (Promega), 1,25mM de $MgCl_2$, 10µL de tampon 5X (Promega), 0.5mM de chaque dNTP et enfin 1µM de chaque amorce selon l'amplifiat désiré. Les séquences des amorces et les conditions d'amplification sont présentées dans le Tableau 13. La Figure 35 (page suivante) illustre la position des amorces sur les gènes étudiés. Les amplifications PCR sont réalisées sur les appareils Mastercycler gradient d'Eppendorf ou iCycler de Biorad. La réaction d'amplification est composée d'une étape de dénaturation initiale à 95°C d'une durée de 5min, suivie de 35 cycles comprenant, une phase de dénaturation à 95°C d'1min, une



<u>Figure 35 :</u> Organisation génétique des cinq contigs (A, B, C, D et E) des gènes codant les enzymes clés des voies de biodégradation des HAP pour la souche *Sphingomonas paucimobilis* sp. EPA505. La taille de chaque gène et les espaces intergéniques sont indiqués ainsi que la position de chaque amorce utilisée pour les amplifications par réaction de PCR. Nom des gènes : *xylX* : sous-unité α de la toluène dioxygénase, *bphC* : dihydroxynaphtalène dioxygénase, *bphA3* : sous-unité de ferrédoxine, *ahdA2c* : petite sous-unité d'oxygénase, *ahdA1c* : grande sous-unité d'oxygénase, *nahD* : 2-hydroxychromène-2-carboxylate isomérase, *phnA1a* : sous-unité α de la dioxygénase initiale, *phnA2a* : sous-unité β de la dioxygénase initiale, *ahdA4* : ferrédoxine réductase, *bphB* : *cis*-dihydrodiol déshydrogénase et *nahE* : hydratase-aldolase

phase d'hybridation des amorces (dépendant du couple d'amorce utilisé**,** voir Tableau 13), et une étape d'élongation à 72°C d'une minute. Une étape d'élongation finale est réalisée à 72°C pendant 7min.

Les produits PCR sont séparés en fonction de leur taille sur gel d'agarose (de 1 à 3 % (w/v)), dans du tampon TBE 1X [pH 8,0] (90mM Tris-borate, 2mM EDTA), pendant 40min à 60min à 90V. Les acides nucléiques sont visualisés sous UV après une coloration au bromure d'éthidium. Les bandes d'ADN correspondantes aux gènes d'intérêt sont excisées du gel, puis éluées grâce au kit QIAquick Gel Extraction Kit (QIAGEN) selon les recommandations du fournisseur, afin d'éviter toute contamination potentielle. Les fragments purifiés sont insérés dans les vecteurs plasmidiques PCR®II-TOPO (Invitrogen) selon les recommandations du fournisseur. La transformation est réalisée par choc thermique à partir de 3µL du produit de ligature mélangé avec 100µL de cellules compétentes (souche bactérienne d'*E.coli* XL1 blue rendue compétente, selon la méthode de Inoue décrit par Sambrook (Sambrook *et al.*, 2001). Après incubation des cellules bactériennes et du mélange de ligature pendant 30min à 4°C, 2min à 42°C et 5min à 4°C, 100µL de SOC sont ajoutés (10mL de SOB, $MgSO_4$ 10mM, glucose 0,4 %). Une nouvelle incubation de 40min à 37°C est réalisée. Un volume de 50µL de chaque aliquot de cellules transformées est étalé sur un milieu gélosé sélectif de LB agar (15g/L d'agar) additionné de 100µg/mL d'ampicilline sur lequel a été au préalablement étalé 40µL d'IPTG (0,8M) et 40µL de X-Gal (20mg/mL). Les boîtes sont incubées une nuit à 37°C. Les plasmides recombinants sont extraits selon le protocole de lyse alcaline décrit par Sambrook (Sambrook *et al.*, 2001). Les clones sont ensuite analysés par restriction enzymatique. Ainsi, 2,5µL de solution plasmidique sont digérés par 1U d'enzyme *EcoRI* pendant 1h à 37°C. Les produits de digestion sont séparés sur gel d'agarose 1 % dans du tampon TBE 1X pendant 1h à 90V et visualisés sous UV après coloration au bromure d'éthidium.

Le séquençage des clones recombinants est réalisé selon la méthode de Sanger (Sanger *et al.*, 1977) par la société MWG en utilisant les amorces Sp6 (ATTTAGGTGACACTATAGAA) ou T7 (TAATACGACTCACTATAGGG). Les séquences brutes sont alors traitées avec le package Staden (Staden, 1996). Les séquences obtenues sont alors comparées aux banques de données Swiss-Prot et TrEMBL à l'aide du programme BLASTx (Altschul *et al.*, 1990).

2.5.2. Suivis d'expression par RT-PCR quantitative

2.5.2.1. Préparation des gammes étalons pour les gènes ciblés

Tableau 14 : Amorces utilisées durant les étapes de transcription inverse et de RT-PCR quantitative des gènes impliqués dans la dégradation des HAP de la souche *Sphingomonas paucimobilis* sp. EPA505.

Gène	*Transcription inverse*	*RT-PCR quantitative*		
	Amorce	Amorce sens	Amorce antisens	Taille (pb) amplifiat
bphC	AAGCCCGAAAGCGACCGAAT	TTTTACGGGCCGCAAGTCGA	GGAACATCGTCCTGACGCAG	120
bphA3	CCTTCCGGCTGATCGATGCA	GCCATGCTGACCGATGGCTA	GTGGCGATGTCGAAGGAACC	77
ahdA2c	CGTCAGTGCGGGTCATAAGG	CTGCAGCAACTCGTGACCGA	CGGTGTAGACGCAGTCCGAA	106
ahdA1c	ATGGCGATGAAGGAGGTTGC	CCCTATCACGCCAGCCTTCT	AATGCCGCTTCTCGCTATCC	100
phnA1a	GACCGGCACTTTCCAGTTGC	CAGATTTGCCACGCCGACAG	TTCGGGATCATGGCAACCGA	204
phnA2a	CGTGACTCTCGCATCGAGGA	GTCGAAGCCTTTGAAGCCGA	TGAAACCATTGCCGTCCTGG	139
ahdA4	TGTTGACCTCGATCCCGGCA	CCAGTGCCGATGGAAAGAGC	CAAACCTGAAGCCGGTCACG	115
nahE	TGATACAGGTCGTGCCGACG	GGATCAACGGCATCCTGAGC	TCGACCAGCGCACGCATGAA	91
bphB	TCGATGCGACATACAGCGTC	GACGTCAACCTGAAGGGGTA	TCGATGCGACATACAGCGTC	139

Les informations de séquences pour la souche *Sphingomonas paucimobilis* sp. EPA505 ont permis la détermination d'amorces (Figure 35 et Tableau 14), permettant d'amplifier la totalité de chaque gène (sauf pour *ahdA4*, *bphB* et *nahE* où seuls des fragments ont été obtenus). Chaque amplifiat acquis selon le protocole décrit précédemment est alors inséré dans les vecteurs plasmidiques PCR®II-TOPO (Invitrogen), selon les recommandations du fournisseur. Les étapes de transformation, d'extraction et purification et de séquençage sont les mêmes que celles décrites ci-dessus.

Les plasmides contenant un fragment ou la totalité des gènes obtenus (*phnA1a, phnA2a, ahdA1c, ahdA2c, bphA3, ahdA4, bphC, bphB* et *nahE*) en utilisant les amorces listées dans le Tableau 13, sont linéarisés par restriction enzymatique afin de libérer l'extrémité 3' du gène. Les enzymes de restriction utilisées sont *Xba*I, *Bam*HI ou *Hind*III (Promega), selon l'orientation de l'insert dans le vecteur PCR®II-TOPO. Le choix de l'enzyme de restriction s'est également appuyé sur l'absence de site de restriction correspondant dans le gène étudié afin de ne pas obtenir plusieurs fragments. Le produit de restriction est vérifié sur gel d'agarose 0,8 %, et purifié sur colonne par le « MinElute PCR Purification kit » (Qiagen) selon les recommandations du fournisseur.

La transcription *in vitro* des gènes est réalisée, suivant l'orientation de l'insert, avec le MEGAscript kit T7 ou Sp6 (Ambion) selon les recommandations du fournisseur mais avec modification du mélange réactionnel : on ajoute en plus 20U de RNasin Inhibitor + (Promega) pour éviter toute dégradation. L'élimination de l'ADN matrice se fait par traitement avec 2 unités (U) de TURBO DNase (Ambion) pendant 15 min à 37°C. Les ARN sont ensuite purifiés sur colonne en utilisant le « RNeasy MinElute Cleanup kit » (Qiagen). La quantité et l'intégrité des ARN sont estimées en utilisant respectivement le Nanodrop spectrophotometer (Nanodrop), et le RNA 6000 Nano kit de l'Agilent 2100 Bioanalyzer (Agilent Technologies), selon les recommandations des fournisseurs.

La réaction de transcription inverse est réalisée à 55°C pendant 2 heures avec 50 ng d'ARN, 100 unités (U) de SuperScriptIII reverse transcriptase (Invitrogen), 1 unité (U) de RNasin Inhibitor +, 0,25 mM de dNTPs (Invitrogen), 0,1 M de dithiothréitol (DTT, Invitrogen), 0,625 µM de chaque amorce du mélange présenté dans le Tableau 14, et 4 µL de tampon Invitrogen FS 5X dans un volume final de 20 µL selon les recommandation du fournisseur. Les ADNc sont ensuite traités 30 min à 37°C avec 2 unités (U) de RNase H (Invitrogen), dénaturés à 95°C pendant 10min et purifiés sur colonne avec le « MinElute PCR Purification kit ». La qualité et la concentration des ADNc obtenus sont déterminées par dosage au Nanodrop (Nanodrop). Chaque ADNc est alors dilué en série (d'un facteur 10) dans

de l'eau nanopure pour générer une gamme s'échelonnant de, $4{,}37.10^7$ copies/µL, à 4,37 copies/µL.

Le nombre de copies de transcrit est calculé en allouant une masse moléculaire moyenne de 303,7 Da par base d'ADNc simple brin, selon la formule suivante :

$$Nombre\ de\ copies\ de\ transcrit(molecules/\mu L) = \frac{quantité\ \ d'ADNc\ \ (g/\mu L)}{Longueur\ \ (bases)\ \times\ 303.7} \times A$$

A = Nombre d'Avogadro = 6.022×10^{23} molécules/mole

2.5.2.2. Quantification de l'expression génique

L'étape de transcription inverse est réalisée comme décrit précédemment, à partir de 50ng d'ARN total en utilisant un mélange de plusieurs amorces ciblant les gènes étudiés (0,625µM pour chaque amorce), afin de minimiser les biais de manipulation : *bphC, ahdA1c, ahdA2c* et *bphA3* pour le premier mélange, *phnA1a, phnA2a* et *ahdA4* pour le deuxième mélange et *bphB* et *nahE* pour le troisième mélange (voir Tableau 14). Les tests réalisés, au préalable, sur l'impact du mélange de ces amorces sur l'efficacité de transcription inverse, ont montré des résultats identiques à ceux obtenus lorsque les amorces étaient utilisées individuellement. Un témoin de contamination d'ADN génomique est réalisé pour chaque échantillon. Chaque réaction de transcription inverse est réalisée en triplicat. Les ADNc obtenus sont ensuite dilués au 1/10 avant d'être analysés en PCR quantitative.

Les expériences en PCR quantitative sont réalisées avec le « MESA Green qPCR for SYBR assays kit » (Eurogentec) selon les instructions du fournisseur. Toutes les amplifications sont réalisées dans un volume final de 20µL contenant 5µL d'échantillon ou 5µL de standard, 10µL de mélange « MESA Green qPCR for SYBR assays » 2X, et une concentration finale de 0,2µM de chaque amorce. Les combinaisons d'amorces utilisées pour chaque gène sont présentées dans le Tableau 14. Pour chaque échantillon, la quantification du produit de transcription inverse par PCR quantitative est réalisée en duplicat. Ainsi, l'étape de transcription inverse étant effectué en triplicat, six mesures au total sont obtenues pour chaque échantillon de la cinétique de bioconversion. Pour les gammes étalon, chaque point de dilution est quantifié en triplicat.

La réaction de PCR est réalisée avec l'appareil Mastercycler Realplex (Eppendorf) selon le programme suivant : une dénaturation initiale de 5min à 95°C, suivie de 40 cycles constitués, d'une étape de dénaturation 15s à 95 °C, et d'une autre de 45s à 68°C (regroupant l'hybridation et l'élongation). A la fin de ces étapes, une courbe de dissociation est réalisée en mesurant l'intensité de fluorescence du SYBR Green lors d'une augmentation progressive de température de 68°C à 95°C durant 20min. Cette étape permet de détecter la formation

éventuelle de dimères d'amorces, et de s'assurer de l'absence d'amplifications aspécifiques. L'analyse des données est réalisée avec le logiciel Realplex v1.5 (Eppendorf). Les résultats issus des expériences en PCR quantitative ont été analysés et validés selon les recommandations décrites par Nolan (Nolan *et al.*, 2006).

3. Caractérisation des échantillons environnementaux

3.1. Extraction d'ADN total

Le protocole est basé sur la méthode développée par Zhou (Zhou *et al.*, 1996). 13,5mL de tampon d'extraction (100mM Tris-HCl [pH 8,0] (Sigma-Aldrich), 100mM Sodium-EDTA [pH 8,0] (Sigma-Aldrich), 10mM Sodium phosphate [pH8,0] (Sigma-Aldrich), 1,5M NaCl (Sigma), 1 % CTAB (Sigma-Aldrich)) et 10µL de protéinase K (10mg/mL) (Invitrogen) sont ajoutés à 5 g de terre. Le mélange est incubé 30min à 37°C sous agitation horizontale (POS-300 (Grant Bio) à 800 rpm). Après ajout de 1,5mL de SDS 20 % (Sigma-Aldrich), l'ensemble est incubé 2h à 65°C au bain-marie SW23 (Julabo), à 110rpm, avec une homogénéisation manuelle toutes les 20min. Après centrifugation (8 000g, 10min, température ambiante, Multifuge 3 SR (Heraeus)), le surnageant est prélevé, et le culot est lavé deux fois par addition de 4,5mL de tampon d'extraction, et 0,5mL de SDS 20 %. A chaque fois, l'ensemble est incubé 10min à 65°C, puis centrifugé 10min à 8 000g pour récupérer le surnageant. L'ADN présent dans les surnageants est ensuite purifié par des extractions successives au phénol (Sigma-Aldrich), au phénol/chloroforme/alcool isoamylique (25:24:1) (Sigma-Aldrich) et au chloroforme/alcool isoamylique (24:1) (Sigma-Aldrich). Après précipitation par 0,7 volume d'isopropanol (Sigma-Aldrich), l'ADN est lavé à l'éthanol 70 %, et repris dans 500µL d'eau distillée stérile. La qualité des acides nucléiques est évaluée par migration électrophorétique sur gel d'agarose 1 %. Les extractions sont réalisées en triplicat et poolées de manière à diminuer les biais d'extraction.

3.2. Amplification PCR, clonage, séquençage des gènes ciblés

3.2.1. Etude du gène codant l'ARNr 16S

Le surnageant contenant l'ADN extrait est récupéré et différentes dilutions de celui-ci sont testées par PCR afin de déterminer la meilleure dilution à utiliser pour une bonne efficacité de PCR. L'amplification du gène codant l'ARNr 16S est réalisée par l'utilisation d'amorces dégénérées provenant de la littérature et démontrées comme universelles (amorce

directe 27F : AGAGTTTGATCMTGGCTCAG et amorce antisens 1492R: TACGGYTACCTTGTTACGGACT) (Lane, 1991; Baker *et al.*, 2003).

La réaction est effectuée dans un volume final de 50µL et contient 0,4µM de chaque amorce (Eurogentec), 5mM de $MgCl_2$, 0,25mM de chaque dNTP, 5µL d'ADN génomique à différentes dilutions (1/10, 1/100, 1/500 et 1/1000), 2 unités (U) de GoTaq DNA polymérase (Promega), et 10µl de tampon 5X. Le programme d'amplification compte 30 cycles comprenant une étape de dénaturation à 95°C de 60s, une étape d'hybridation de 30s à une température 59°C et une étape d'élongation de 60s à 72°C. Le thermocycleur utilisé est un iCycler (Biorad). Ces cycles sont précédés d'une étape de dénaturation de 5min à 95°C. Le programme se termine par une étape d'élongation de 7min à 72°C.

Les produits PCR sont séparés en fonction de leur taille sur gel d'agarose 1,2 % (w/v) dans du tampon TBE 1X [pH 8,0] (90mM Tris-borate, 2mM EDTA) pendant 60min à 90V. Les bandes obtenues, et à la bonne taille sont alors, excisées et purifiées comme décrit précédemment. Chaque amplifiat est alors inséré dans les vecteurs plasmidiques PCR®II-TOPO (Invitrogen) selon les recommandations du fournisseur. Les étapes de transformation, d'extraction et purification et de séquençage sont les mêmes que celles décrites ci-dessus (voir paragraphe ***Amplification PCR, clonage, séquençage des gènes ciblés***, page 86).

Les séquences brutes sont tout d'abord traitées avec le package Staden (Staden, 1996). Ces traitements permettent par exemple d'éliminer les séquences de mauvaise qualité et de créer des contigs. Les programmes de ce package permettent au final d'obtenir des séquences consensus de qualité qui sont sauvegardées au format FASTA. Puis, les séquences chimériques sont éliminées par l'outil MALLARD (Ashelford *et al.*, 2006). Les séquences de la banque ainsi obtenue sont tout d'abord comparées à la base de données du site NCBI (http://www.ncbi.nlm.nih.gov/) avec le logiciel BLASTn, les séquences de référence les plus proches étant récupérées. L'alignement de toutes ces séquences est ensuite réalisé via l'outil MEGA4 pour Windows (Tamura *et al.*, 2007). L'arbre phylogénétique est construit par la méthode du plus proche voisin (Saitou et Nei, 1987), permettant l'affiliation des séquences obtenues. La vérification de la robustesse de la construction a été effectuée par un « bootstrap » sur 1 000 constructions. Enfin, les programmes DOTUR (Schloss et Handelsman, 2005), SPADE (http://chao.stat.nthu.edu.tw/softwareCE.html) et PAST (Hammer *et al.*, 2001) ont été utilisés pour calculer les estimations de composition des communautés, pour définir la courbe de raréfaction, et pour calculer les indices de diversité (CGoods et Chao).

3.2.2. Etude du gène codant l'enzyme PhnA1a

L'amplification du gène codant l'enzyme PhnA1a est effectuée selon le même principe que pour le gène codant l'ARNr 16S. L'ADN extrait dilué (1/10, 1/50, et 1/100) est utilisé comme matrice pour l'amplification PCR L'amplification du gène est réalisée par l'utilisation d'amorces dégénérées définies manuellement (amorce directe A1f_2dI : TACCATGTIGGATGGAC et amorce antisens A1f_ 3dIR8: CATGTTITCICYGTCITC).

La réaction est effectuée dans un volume final de 25µL et contient 0,1µM de chaque amorce (Eurogentec), 5mM de $MgCl_2$, 0,25mM de chaque dNTP, 5µL d'ADN génomique à différentes dilutions (1/10, 1/100, 1/500 et 1/1000), 2 unités (U) de GoTaq DNA polymérase (Promega), et 10µl de tampon 5X. L'ajout de 5µL de diméthyl sulfoxide (DMSO, ICN Biomedicals) dans le mélange réactionnel permet d'améliorer la dénaturation de l'ADN et de réduire la formation de structures secondaires pouvant perturber l'action de la polymérase. Le programme d'amplification compte 40 cycles comprenant, une étape de dénaturation à 95°C de 30s, une étape d'hybridation de 30s à une température de 46°C, et une étape d'élongation de 60s à 72°C. Le thermocycleur utilisé est un iCycler (Biorad). Ces cycles sont précédés d'une étape de dénaturation de 5min à 95°C. Le programme se termine par une étape d'élongation de 7min à 72°C.

Les produits PCR sont séparés en fonction de leur taille sur gel d'agarose 1,2 % (w/v) dans du tampon TBE 1X [pH 8,0] (90mM Tris-borate, 2mM EDTA) pendant 60min à 90V. Les bandes attendues sont alors excisées et purifiées comme décrit précédemment. Chaque amplifiat est alors inséré dans les vecteurs plasmidiques PCR®II-TOPO (Invitrogen) selon les recommandations du fournisseur. Les étapes de transformation, d'extraction et purification et de séquençage sont les mêmes que celles décrites ci-dessus (voir paragraphe ***Amplification PCR, clonage, séquençage des gènes ciblés***, page 86).

4. Biopuces ADN

4.1. Préparation des échantillons et marquage

4.1.1. Echantillons ADN

L'ADN environnemental extrait de la terre est amplifié et marqué en utilisant l'Alexa Fluor® 3 ou 5 via l'application du kit « BioPrime® Array CGH Genomic Labeling System » (Invitrogen) selon les recommandations du fournisseur. La réaction est effectuée en triplicat afin de diminuer les biais d'amplification. Les produits obtenus sont enfin fragmentés par une étape à 95°C pendant 15 minutes dans l'appareil iCycler de Biorad.

4.1.2. Echantillons ARN

L'ARN total de la souche *Sphingomonas paucimobilis* sp. EPA505 extrait à différents temps des cinétiques de croissance (3, 6, 10 et 21h) a tout d'abord été enrichi avec le kit « MICROB*Express*™ Bacterial mRNA Enrichment » (Ambion) selon les recommandations du fournisseur. Selon les échantillons, de 500ng à 2µg ont été traités.

Les ARNm enrichis ont alors été amplifiés en ARNa (ou ARN antisens) avec le kit « MessageAmp™ II-Bacteria RNA Amplification » (Ambion) selon les recommandations du fournisseur. Toutefois, l'étape de transcription *in vitro* a été modifiée pour pouvoir incorporer de l'amino-allyl-UTP ou aaUTP (Ambion) pour l'étape de marquage ultérieure. La réaction est effectuée dans 40µL total contenant : l'échantillon ADN double-brin purifié (14µL environ), 4µL de tampon 10X (Ambion), 4µL d'enzyme de transcription *in vitro* (Ambion) 7,5mM d'ATP, de CTP et de GTP (Ambion), 5,6mM d'UTP (Ambion), 3.76mM d'aaUTP (Ambion). Le mélange réactionnel est incubé à 37°C pendant 14h, dans un thermocycleur Mastercycler gradient (Eppendorf). La purification est réalisée comme recommandé par le fournisseur (Ambion).

L'étape de marquage est réalisée sur 10µg d'ARNa. L'échantillon est tout d'abord séché à vide à l'aide d'un évaporateur Vacufuge® SpeedVac (Eppendorf) puis repris dans 10µL de tampon sodium bicarbonate 0,1M [pH 8,7]. Le marquage est alors effectué en utilisant le kit « Amersham CyDye™ Post-Labeling Reactive Dye Packs » (GE Healthcare) en suivant les recommandations du fournisseur. Les échantillons sont marqués en Cyanine3 pour les cultures en présence de phénanthrène et de glucose, et en Cyanine5 pour celles effectuées en présence de fluoranthène ou du mélange de deux polluants. Les échantillons marqués sont ensuite purifiés par « NucleoSpin RNA Clean-Up » (Macherey-Nagel) et repris dans un volume final de 40µL d'H_2O.

Après chaque étape, des dosages qualitatifs et quantitatifs sont effectués. La qualité des ARN extraits est estimée à l'aide du kit RNA 6000 Nano (Agilent Technologies), et à l'appareillage associé (Bioanalyzer 2100). La concentration en acides nucléiques est déterminée via un dosage spectrophotométrique (Nanodrop).

4.2. Caractéristiques des biopuces utilisées, réactions d'hybridation et acquisition des images

4.2.1. Biopuce taxonomique

Une biopuce ADN de 14 399 sondes (d'une taille de 25-mers) ciblant le biomarqueur 16S a été développée. La détermination des sondes a été effectuée en utilisant le logiciel

PhylArray (Militon *et al.*, 2007) développé par l'équipe. La biopuce fabriquée par la société Agilent cible 1 938 genres procaryotiques (1 845 bactéries, et 93 archées). Toutes les sondes présentent en triplicat ont été réparties sur la biopuce de manière aléatoire pour minimiser les variations spatiales potentielles durant l'étape d'hybridation.

Avant hybridation, l'échantillon précédemment marqué en Alexa Fluor® 3 est tout d'abord concentré par évaporation (Vacufuge® SpeedVac ; Eppendorf), puis repris dans un volume final de 11µL d'H_2O. La réaction d'hybridation est effectuée selon le protocole décrit par le fournisseur dans un four à hybridation (Agilent) à 65°C pendant 40 h. Les lavages sont également réalisés comme décrit par le fournisseur. Enfin, les images de chaque plex sont acquises à l'aide du scanner Agilent et des logiciels associés selon les recommandations du fournisseur.

4.2.2. Biopuce fonctionnelle

Une biopuce ADN de 39 216 sondes ciblant les voies de dégradation des HAP a été développée. Les sondes ont été déterminées avec le logiciel Metabolic Design présenté dans ces travaux de thèse. La fabrication de la biopuce a été réalisée par l'entreprise Roche Nimblegen. Des espaceurs (dans ce cas des thymines), ont également été rajoutés en position 3', pour que chaque sonde fasse au minimum 26-mers. Toutes les sondes présentent en triplicat ont été réparties sur la biopuce de manière aléatoire, pour minimiser les variations spatiales potentielles, durant l'étape d'hybridation. Des oligonucléotides contrôles appelés « RANDOM » (8 863 sondes avec des tailles aléatoires allant de 20 à 56-mers), sont également présents sur la biopuce et servent à la détermination du bruit de fond.

Les échantillons à hybrider sont tout d'abord concentrés par évaporation (Vacufuge® SpeedVac ; Eppendorf), puis repris dans un volume final de 5,6µL d'H_2O. La réaction d'hybridation est effectuée selon le protocole décrit par le fournisseur au sein d'un système d'hybridation à quatre chambres (Roche Nimblegen) à 42°C pendant 72 h. Les lavages sont réalisés avec les solutions de lavages I, II et III comme décrit par le fournisseur.

Les images de chaque plex sont acquises indépendamment à l'aide d'un scanner Innoscan 900AL (Innopsys) associé au logiciel Mapix® (Innopsys). Les images analysées ont toutes les mêmes caractéristiques (gain du photomultiplicateur : 30db, vitesse de balayage : 20pixels/s, puissance du laser : normale).

4.3. Extraction des données brutes, prétraitement et analyse des images de biopuces ADN

4.3.1. Biopuce taxonomique

Pour chaque image obtenue, l'extraction des données brutes d'intensité est réalisée en utilisant le logiciel dédié (Agilent's Feature Extraction 10.7 Image Analysis Software), développé par Agilent. Pour cette biopuce, le logiciel dédié détermine automatiquement le bruit de fond local de chacun des spots de la lame qui servira à effectuer les calculs. Pour calculer ce bruit de fond, le logiciel utilise la méthode dite du rayon pour chacun des spots (Source : Feature Extraction v9.5 Reference Guide d'Agilent, Chapitre 5 : « How each algorithm calculates a result »). Les signaux de chacun des spots sont ensuite analysés, et seuls les spots donnant un signal cinq fois supérieur au bruit de fond local sont considérés comme positifs. Enfin, pour qu'un genre soit considéré comme présent, plus de 50 % des sondes ciblant ce genre particulier doivent répondre (le nombre de sondes varie entre 1 à 4 selon le nombre de régions ciblées).

4.3.2. Biopuce fonctionnelle

Pour chaque image obtenue, l'extraction des données brutes d'intensité est réalisée en utilisant le logiciel NimbleScan v2.1. (Roche NimbleGen).

Le bruit de fond est calculé en utilisant les données d'intensité de signal des sondes RANDOM présentes au sein de chaque lame. La méthode de calcul consiste tout d'abord en un découpage de la lame en seize zones, qui seront analysées indépendamment (pour minimiser la variation spatiale du signal potentielle). Pour chacune de ces seize zones, le bruit de fond est alors défini par deux valeurs : l'intensité médiane (Bposition), et sa dispersion (Bdispersion). Un rapport signal/bruit nommé SNR' est ensuite calculé pour chaque sonde, pour la zone considérée, par la formule (Verdick *et al.*, 2002) :

$$SNR' = (Signal - Bposition) / Bdispersion$$

Ce calcul permet ainsi de centrer et de réduire les données brutes d'intensité pour chaque sonde et donc d'éliminer potentiellement les variations spatiales de signal au sein d'une même image. La valeur Bdispersion est déterminée selon la « propreté » de l'image obtenue en se basant sur le coefficient de variation (C_v) des sondes RANDOM pour chaque zone (il est déterminé en calculant le rapport entre l'écart-type et la moyenne des signaux des sondes RANDOM). Si C_v a une valeur supérieure à 33,33 %, l'image est dite « sale », et Bdispersion sera calculée selon la formule :

$$Bdispersion = Bposition - Q3$$

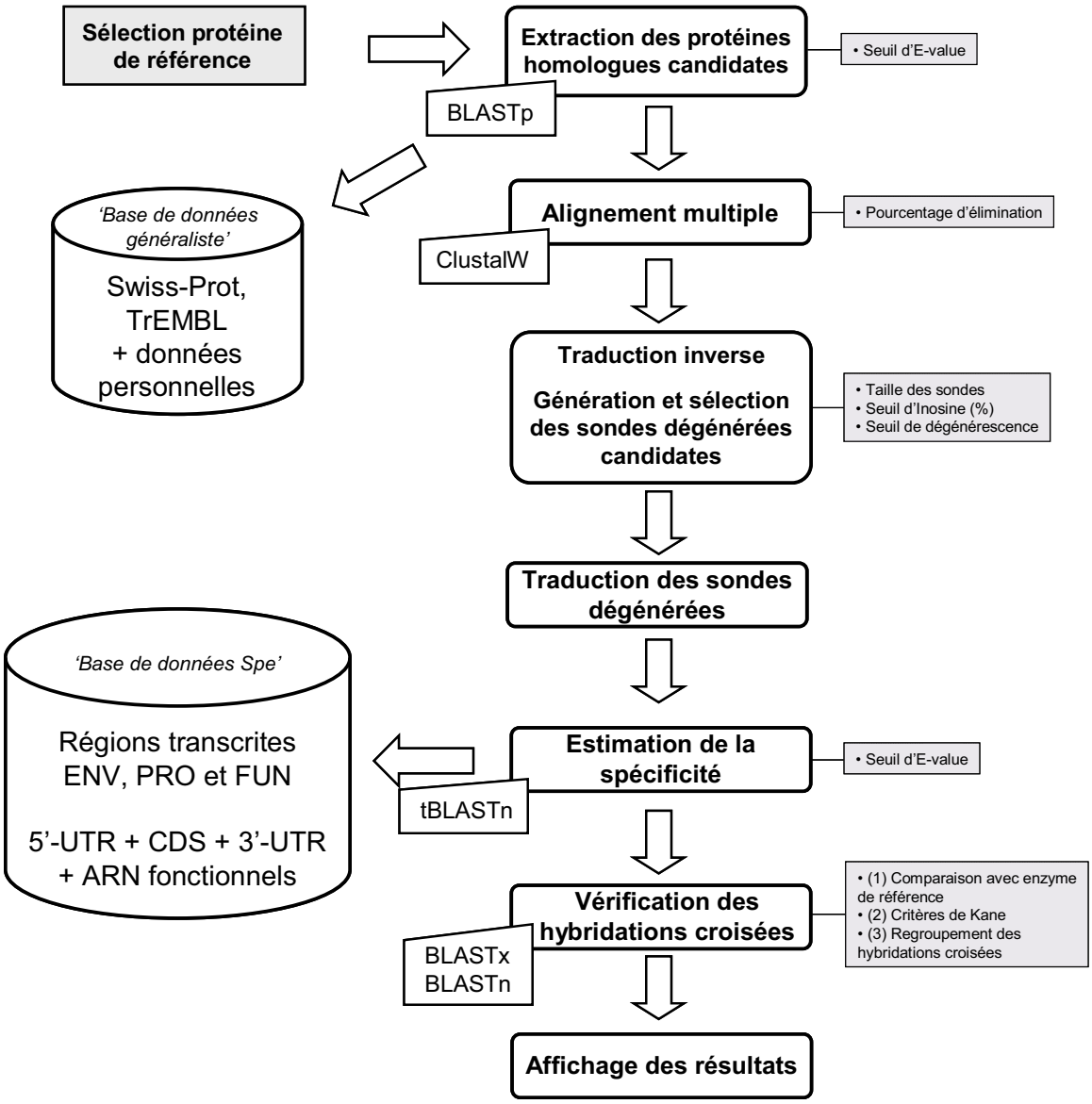
où Q3 est le troisième quartile (en statistique descriptive, un quartile est chacune des 3 valeurs qui divisent un jeu de donnée en 4 parts égales, triées selon une relation d'ordre, de sorte que chaque partie représente un quart de l'échantillon de population total) calculé sur les intensités des sondes RANDOM.

Si C_V a une valeur inférieure à 33,33 %, l'image est dite « propre », et Bdispersion sera calculée selon la formule (où D8 est le septième décile (en statistique descriptive, un décile est chacune des 9 valeurs qui divisent un jeu de donnée en 10 parts égales, triées selon une relation d'ordre, de sorte que chaque partie représente un dixième de l'échantillon de population total) calculé sur les intensités des sondes RANDOM) :

$$\boxed{Bdispersion = Bposition - D8}$$

Un script en langage PERL permet d'effectuer automatiquement cette analyse et ces calculs sur les données brutes de chaque image. Un second script permet ensuite de regrouper les réplicats de chaque sonde et de calculer pour chacune la médiane et l'écart type obtenu. Les sondes dites positives sont celles ayant un SNR' supérieur à 3 (valeur limite permettant d'éviter les faux positifs) (He et Zhou, 2008).

RESULTATS



<u>Figure 36 :</u> Stratégie implémentée dans le programme Metabolic Design pour déterminer des sondes exploratoires à partir d'une séquence protéique de référence. Pour chaque étape, un ou plusieurs paramètres peuvent être définis par l'utilisateur. La première étape consiste en l'extraction des protéines homologues via une analyse BLASTp contre la base de données regroupant les bases Swiss-Prot et TrEMBL. La seconde étape permet : d'éliminer les séquences trop divergentes en taille par rapport à la séquence de référence (pourcentage d'élimination), et de réaliser en local l'alignement multiple des séquences choisies et de la séquence de référence. La troisième étape permet la génération et la sélection des sondes dégénérées candidates à partir de la séquence nucléique dégénérée consensus déduite de l'alignement multiple protéique. La spécificité des sondes « exploratoires » *in silico* est évaluée après traduction des sondes dégénérées (formant ainsi de nouvelles combinaisons) en utilisant la base de données « Base de Données Spe ». La vérification des hybridations croisées s'effectue en trois grandes étapes : (1) en comparant la séquence du gène engendrant une hybridation potentielle à celle de la protéine de référence par BLASTx ; (2) si la séquence est trop éloignée, Metabolic Design évalue les critères de Kane entre la séquence du gène et celle de la sonde par BLASTn ; (3) enfin, toutes les séquences engendrant potentiellement des hybridations croisées pour chaque sonde dégénérée sont regroupées en « familles » par BLASTn, une seule hybridation croisée par famille est décomptée. Les résultats sont organisés et affichés dans une page HTML.

Chapitre I : Détermination de sondes pour biopuce métabolique dite exploratoire

1. Introduction

L'objectif est de définir des sondes ciblant les gènes codant des enzymes impliquées dans les voies métaboliques de dégradation de divers HAP et d'autres composés aromatiques. A l'heure actuelle, les outils existants de détermination de sondes pour biopuces ADN ne permettent que de cibler les gènes dont les séquences ont été déterminées. Or, malgré la multiplication des projets de séquençage à grande échelle, et l'évolution des techniques de séquençage, il est actuellement impossible de connaître toute la diversité génétique de chaque gène. De ce fait, pour étudier les potentialités métaboliques d'un écosystème donné, sans *a priori* sur les séquences géniques ciblées, il est nécessaire de développer des sondes dites « exploratoires », c'est-à-dire capables d'appréhender toute la diversité génique.

L'outil permettant la conception de telles sondes doit être simple d'utilisation, et facile d'accès. Préalablement à l'élaboration des sondes, l'outil doit d'abord permettre l'extraction des données de séquences qui seront utilisées. Plusieurs gènes pouvant être étudiés en parallèle, il est judicieux de développer un outil d'extraction travaillant à partir d'une visualisation graphique des étapes métaboliques étudiées. L'utilisateur doit ainsi pouvoir élaborer, à façon, cette visualisation offrant ainsi une très grande flexibilité du système. Pour s'affranchir des biais du code génétique, enfin, l'extraction des données se fera sur les séquences protéiques.

2. Conception du logiciel Metabolic Design

2.1. Approche globale

L'outil, appelé Metabolic Design, qui a été développé permet donc : (i) de définir à façon et de visualiser une voie métabolique ; (ii) de réaliser simultanément, pour chacune des enzymes d'intérêt, une fouille des données de génomique, afin d'identifier l'ensemble des séquences présentant des similarités ; (iii) et de définir en utilisant les séquences présélectionnées par la fouille de données, des sondes pour biopuces ADN fonctionnelles dites « exploratoires », spécifiques de chaque étape métabolique (Figure 36).

```
>sp|P11122|BPHC_PSEPA Biphenyl-2,3-diol 1,2-dioxygenase OS=Pseudomonas
paucimobilis GN=bphC PE=3 SV=1
MVAVTELGYLGLTVTNLDAWRSYAAEVAGMEIVDEGEGDRLYLRMDQWHHRIVLHASDSD
DLAYLGWRVADPVEFDAMVAKLTAAGISLTVASEAEARERRVLGLAKLADPGGNPTEIFY
GPQVDTHKPFHPGRPMYGKFVTGSEGIGHCILRQDDVPAAAAFYGLLGLRGSVEYHLQLP
NGMVAQPYFMHCNERQHSVAFGLGPMEKRINHLMFEYTDLDDLGLAHDIVRARKIDVALQ
LGKHANDQALTFYCANPSGWLWEFGWGARKAPSQQEYYTRDIFGHGNEAAGYGMDIPLG
```

<u>Figure 37 :</u> Exemple de séquence au format FASTA issue de la base de données de haute qualité Swiss-Prot.

La ligne de commentaire est organisée de la manière suivante :

>Libellé de la base|Numéro d'accession|Nom d'entrée Nom de la protéine OS=Nom de l'organisme GN=Nom du gène PE=Existence protéique SV=Version de la séquence.

Le libellé de la base indique l'origine de la séquence (sp : Swiss-Prot, tr : TrEMBL), puis vient le numéro d'accession et le nom d'entrée (couplant le nom du gène et le nom de l'organisme source). Vient ensuite le nom de la protéine (ici : 'Biphenyl-2,3-diol 1,2-dioxygenase'), le nom de l'organisme précédé de 'OS=' et le nom du gène codant pour la protéine précédé de 'GN='. Enfin, il est précisé le nombre d'ADNc (indiquée par 'PE='), disponibles dans les bases correspondant à la séquence protéique et le numéro de version de la séquence enregistrée au sein de la banque de référence (indiquée par 'SV=').

Pour assurer la fouille de données, une séquence protéique, dite de référence, est recherchée dans une première base de données de haute qualité. Cette fouille est effectuée en termes d'annotation fonctionnelle, par la recherche de mots-clés ou de numéros d'accessions fournis par l'utilisateur. La séquence de référence choisie est ensuite utilisée comme séquence requête par le logiciel BLASTp. Cette étape permet d'extraire l'ensemble des séquences qui lui sont similaires au sein des bases de données internationales et/ou de données personnelles. Après avoir sélectionné l'ensemble, ou une partie des séquences extraites, Metabolic Design réalise un premier filtre en éliminant les séquences trop divergentes en taille de la séquence de référence (en utilisant le pourcentage d'élimination défini par l'utilisateur). Les séquences restantes sont ensuite alignées avec l'outil ClustalW. Cet alignement multiple est le point de départ du module visant à déterminer les sondes oligonucléotidiques spécifiques de chaque enzyme. Ces sondes sont définies à partir de la séquence nucléotidique consensus dégénérée, déduite de l'alignement protéique multiple, après traduction inverse. Ces sondes, pour être valides, doivent aussi répondre à plusieurs critères définis par l'utilisateur (taille, dégénérescence, pourcentage d'Inosine). Enfin, la spécificité de chaque sonde est vérifiée en recherchant les hybridations croisées potentielles. Cette recherche est effectuée en comparant les sondes avec une base de données répertoriant l'ensemble des séquences potentiellement transcrites au sein des écosystèmes étudiés. Pour chaque séquence de la base présentant une similarité avec les sondes, plusieurs tests (qui seront détaillés plus loin) vérifient si elles peuvent engendrer des hybridations croisées. Par exemple, les critères de Kane sont vérifiés pour chaque hybridation croisée (Kane *et al.*, 2000). Metabolic Design a été implémenté en différents modules pour une meilleure maintenance, une réutilisation de ces derniers de manière indépendante, et enfin pour faciliter les évolutions futures.

2.2. Mise en forme des bases de données, réorganisation des informations

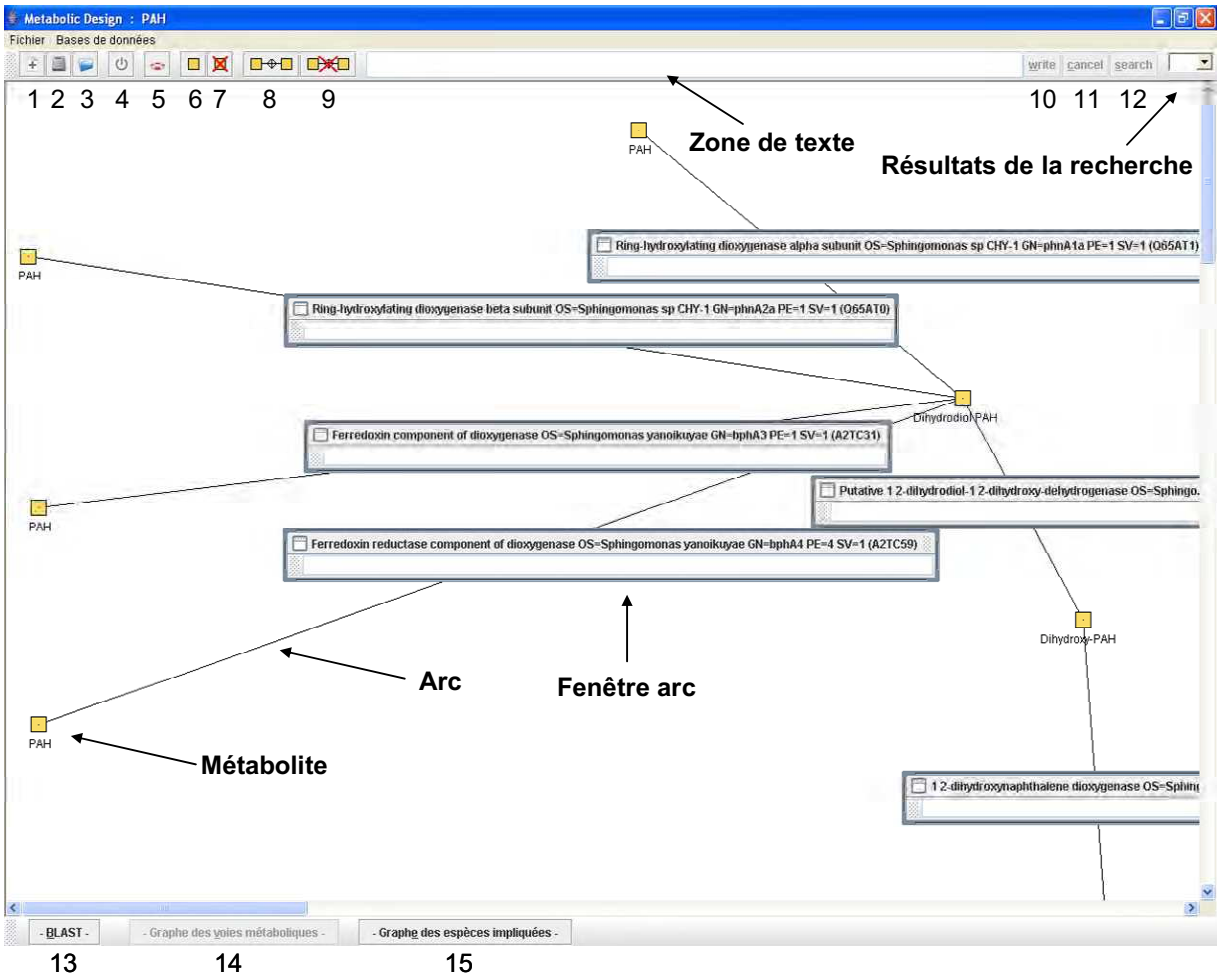
Au cours des différentes étapes de fouille de données et des tests de spécificité des sondes, le logiciel Metabolic Design fait appel à différentes banques de données construites à partir des bases de données internationales. Chacune de ces bases peut également intégrer des données propriétaires.

(1) La base de données de haute qualité a été développée a partir des données au format FASTA de la banque Swiss-Prot (voir Figure 37 pour plus de détails sur les séquences au format FASTA). Cette dernière regroupe des séquences protéiques dont les informations sont contrôlées manuellement. Elle s'efforce ainsi de fournir des séquences

de protéines fiables, avec des annotations fonctionnelles de qualité validées le plus souvent expérimentalement. Elle est utilisée pour définir les séquences de référence utilisées pour faire la fouille de données.

(2) La base de données protéique à partir de laquelle est réalisée la fouille des données regroupe toutes les données UNIPROT. Ainsi, cette base contient ainsi à la fois des séquences de Swiss-Prot, mais également celles de la base TrEMBL. Les données d'UNIPROT, importées au format EMBL, sont tout d'abord traitées, afin de ne récupérer que les fiches reliées aux domaines des *Bacteria*, *Archaea* et *Fungi*. Puis, les données de ces fiches sont extraites et concaténées en un seul fichier, pour former la base de données protéique. De plus, afin de faciliter l'accès aux données via l'interface graphique, ces données sont également sauvegardées dans deux dossiers. Le premier contient les séquences protéiques au format FASTA, chacune d'entre elle étant stockée séparément dans un fichier ayant pour nom le numéro d'accession de la séquence protéique. Il est important de noter que tous les fichiers séquences provenant d'une même espèce sont localisés dans un répertoire portant le nom de l'espèce considérée. Le deuxième dossier possède la même structure, mais contient les données au format EMBL. Cette organisation permet un accès rapide aux informations que l'utilisateur demande pour l'affichage à l'écran.

(3) La base de données (appelée « Base de Données Spe »), définie pour évaluer *in silico* les hybridations croisées potentielles, est construite à partir des données de la base EMBL. Cette banque contient uniquement les séquences nucléiques potentiellement transcrites des divisions procaryotes (PRO), champignons (FUN) et environnement (ENV) de la base EMBL. En effet, la finalité est d'étudier, par l'approche biopuce ADN, les gènes exprimés au sein des écosystèmes en utilisant donc comme cibles les ARNm. A partir des séquences des trois divisions sélectionnées, tous les CDS ont donc été récupérés, ainsi que les régions UTR 3' et 5'. Lorsque les informations sur les bornes des UTR étaient absentes, 100 nucléotides de part et d'autre des CDS ont été extraits arbitrairement. Enfin, les séquences des ARN fonctionnels (ARNt, ARNr,…) ont également été intégrées à cette base, car ces molécules sont présentes dans les ARN totaux extraits des échantillons environnementaux.



<u>Figure 38 :</u> Interface graphique de Metabolic Design pour la visualisation des voies métaboliques et la recherche des séquences de référence.

Les carrés jaunes symbolisent les métabolites liés potentiellement entre eux par des arcs sur lesquelles les enzymes sont représentées sous forme de fenêtre. Pour définir les arcs, il est tout d'abord nécessaire de choisir une enzyme de référence. Pour cela, la zone de texte permet de réaliser les recherches d'enzyme de référence par mots-clés ou numéros d'accession , les résultats sont ensuite affichés dans la liste de sélection en haut à droite de la fenêtre. Après sélection d'une séquence, l'arc est crée par l'utilisateur.

Boutons et options disponibles au sein de l'interface de Metabolic Design : (1) Nouveau fichier, (2) Enregistrer, (3) Ouvrir, (4) Quitter, (5) Aide, (6) Nouveau métabolite, (7) Supprimer métabolite sélectionné, (8) Nouvel arc, (9) Supprimer arc sélectionné, (10) Valider la création d'un arc ou d'un métabolite (11) Annuler et (12) Rechercher (utilisé durant la recherche avec les mots-clés, (13) Lancement du BLAST, (14) Sélectionner l'interface de création des voies métaboliques (reconstruction et édition *in silico* uniquement), (15) Sélectionner l'interface de visualisation des résultats de la fouille des données (affichage des résultats après l'étape de la fouille de données).

2.3. Module de visualisation et de reconstruction des voies métaboliques

Ce module, écrit en JAVA pour l'interface graphique, et en PERL pour le traitement et la mise en forme des données, a été implémenté par Fabrice Gravelat durant son stage d'ingénieur ISIMA au sein de l'équipe de recherche.

Le premier module de Metabolic Design permet la reconstruction *in silico* et la visualisation des différentes étapes métaboliques ciblées d'un processus biologique d'intérêt. C'est avec ce double objectif que l'interface graphique de Metabolic Design a donc été développée (Figure 38). Après l'étape de reconstruction, les métabolites sont matérialisés par des carrés jaunes, reliés entre eux par une ou plusieurs fenêtres, nommés arcs. Ces arcs représentent les enzymes impliquées dans les différentes réactions enzymatiques. Au sein de l'interface de création et d'édition, une ou plusieurs voies métaboliques peuvent être reconstruites via l'utilisation des boutons disponibles et de la zone de saisie.

Pour chaque étape enzymatique de la voie métabolique à construire, l'utilisateur doit tout d'abord choisir l'enzyme impliquée (par la recherche de mots-clés ou de numéro d'accession). Il est possible d'exclure un mot-clé durant la recherche, en précédant le mot du caractère '!'. Les correspondances trouvées dans la base de données de haute qualité apparaissent alors dans une autre zone de saisie pour que l'utilisateur puisse effectuer un choix. Après avoir choisi la séquence, l'utilisateur peut créer l'arc considéré entre deux métabolites en utilisant le bouton 'write' (Figure 38). Cet outil laisse une totale liberté à l'utilisateur, permettant ainsi de relier plusieurs enzymes à un même substrat, ou inversement. Il est donc possible de prendre en compte des enzymes multimériques ou des complexes enzymatiques. De plus, si aucune correspondance n'est trouvée, ou que l'enzyme est inconnue, l'utilisateur peut tout de même créer l'arc désiré en entrant le caractère 'X' (le programme ne tiendra pas compte de cet arc pour les analyses futures).

2.4. Fouille des données, réorganisation et visualisation des résultats

La fouille des données de la base de données généraliste s'appuie sur une recherche de similarité de séquences avec le logiciel BLASTp (Altschul *et al.*, 1990). Les séquences protéiques de référence choisies sont utilisées comme séquences requêtes, et l'utilisateur défini la valeur maximale de l'E-value pour la recherche. Pour chaque enzyme de référence, le fichier de sortie BLASTp est filtré, et traité, pour un affichage des données rapide et clair à l'aide du système de gestion de bases de données relationnelles ORACLE. En effet, pour chaque graphique (comportant une ou plusieurs voies métaboliques), une table est créée, et chaque donnée est définie par une clé primaire composée de deux éléments : le numéro

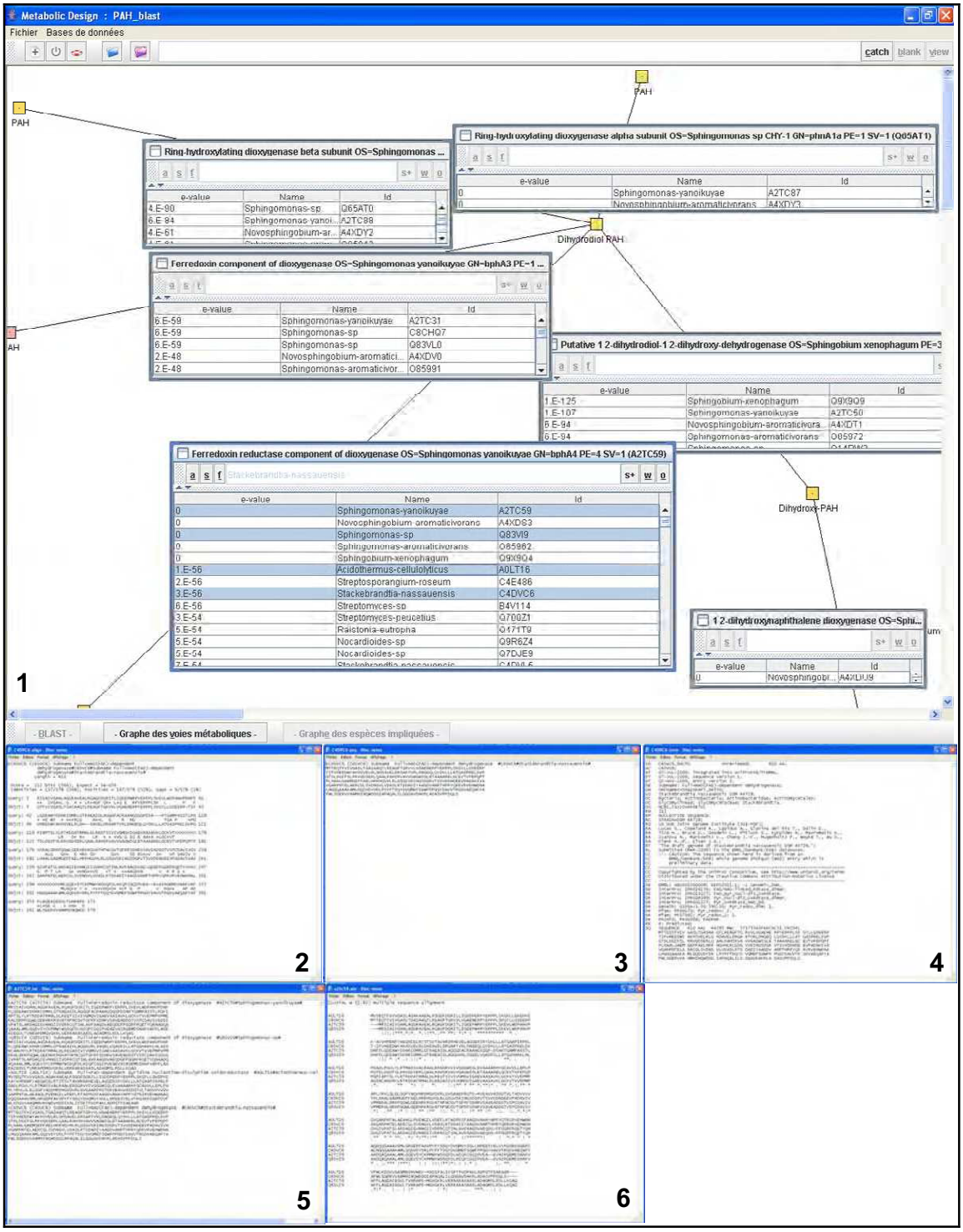


Figure 39 : Interface graphique de Metabolic Design pour la visualisation des résultats de la fouille des données.

(1) Fenêtre principale. La voie précédemment définie est toujours visible à l'écran. Après la fouille de données, les fenêtres internes regroupent les résultats classés selon le la provenance de la séquence (nom de l'organisme) ou l'identité avec la séquence de référence (E-value). Les boutons « catch », « blank » et « view » sont utilisables pour afficher l'ensemble des enzymes d'une voie métabolique donnée pouvant être retrouvée chez un organisme sélectionné par l'utilisateur. Pour chaque protéine similaire extraite il est possible d'obtenir (2) un alignement via l'outil BLASTp avec la protéine de référence (bouton a) ; (3) l'affichage de la séquence au format FASTA (bouton s) ; (4) l'affichage de la séquence au

d'accession de la protéine de référence, et celui de chaque protéine similaire retrouvée. Cette table contient également la valeur de l'E-value du résultat BLASTp et le nom de l'organisme d'où est issue la séquence de la protéine similaire. Pour chaque enzyme de référence, les résultats obtenus ainsi réorganisés peuvent être affichés (E-value, nom de l'organisme source et numéro d'accession de la séquence de la protéine similaire) (Figure 39). L'affichage est directement effectué dans les fenêtres internes reliant les différents métabolites. Le classement des résultats est laissé au choix de l'utilisateur : soit ces derniers sont classés par ordre alphabétique, selon le nom de l'organisme, soit par similarité de séquence, en s'appuyant sur les valeurs d'E-value.

Les données de sortie BLASTp sont également fragmentées par un script PERL, pour en assurer un accès rapide avec l'interface graphique (Figure 39). Ces données fragmentées, pour chaque protéine similaire, sont stockées dans des fichiers dédiés (nommé selon le numéro d'accession de la protéine similaire). Ces fichiers sont eux-mêmes conservés dans un dossier, nommé par le numéro d'accession de la protéine de référence. Cela permet d'afficher immédiatement l'alignement obtenu suite à la recherche de similarité BLASTp entre la protéine de référence et la protéine similaire choisie, et ce par l'intermédiaire d'un simple bouton (bouton a). L'affichage de ce résultat est réalisé dans une nouvelle fenêtre de l'interface graphique (Figure 39).

Il est également possible d'afficher diverses données pour chaque protéine similaire, comme la séquence au format EMBL (bouton f), ou au format FASTA (bouton s) (Figure 39). Ces informations sont directement extraites de la base de données réorganisée à partir de laquelle se fait la fouille des données (voir paragraphe **Mise en forme des bases de données, réorganisation des informations**, page 99). Finalement, une dernière option d'affichage permet d'estimer si un organisme donné possède les enzymes nécessaires pour assurer la ou les voies métaboliques ciblées. Pour cela, l'utilisation du bouton « catch » permet de sélectionner l'organisme, et le bouton « view » montre ensuite la capacité de l'organisme à pouvoir réaliser telle ou telle voie métabolique.

La sélection de plusieurs protéines similaires entraîne l'activation de plusieurs autres fonctionnalités : (i) ainsi, il est possible de concaténer dans un seul fichier toutes les séquences protéiques FASTA choisies (via le bouton s+). (ii) Il est également possible de réaliser sur ces séquences un alignement multiple via l'utilisation du programme ClustalW (bouton w).

Enfin, après sélection des protéines similaires désirées dans la fenêtre interne, un simple clic sur le bouton o entraîne l'ouverture d'une nouvelle fenêtre. Cette dernière permet

format EMBL (bouton f). La sélection de plusieurs séquences protéines dans une même fenêtre interne permet également : (5) la réalisation d'un alignement multiple via l'utilisation du programme ClustalW (bouton w) ; (6) la concaténation dans un seul fichier texte de toutes les séquences protéiques au format FASTA (bouton s+).

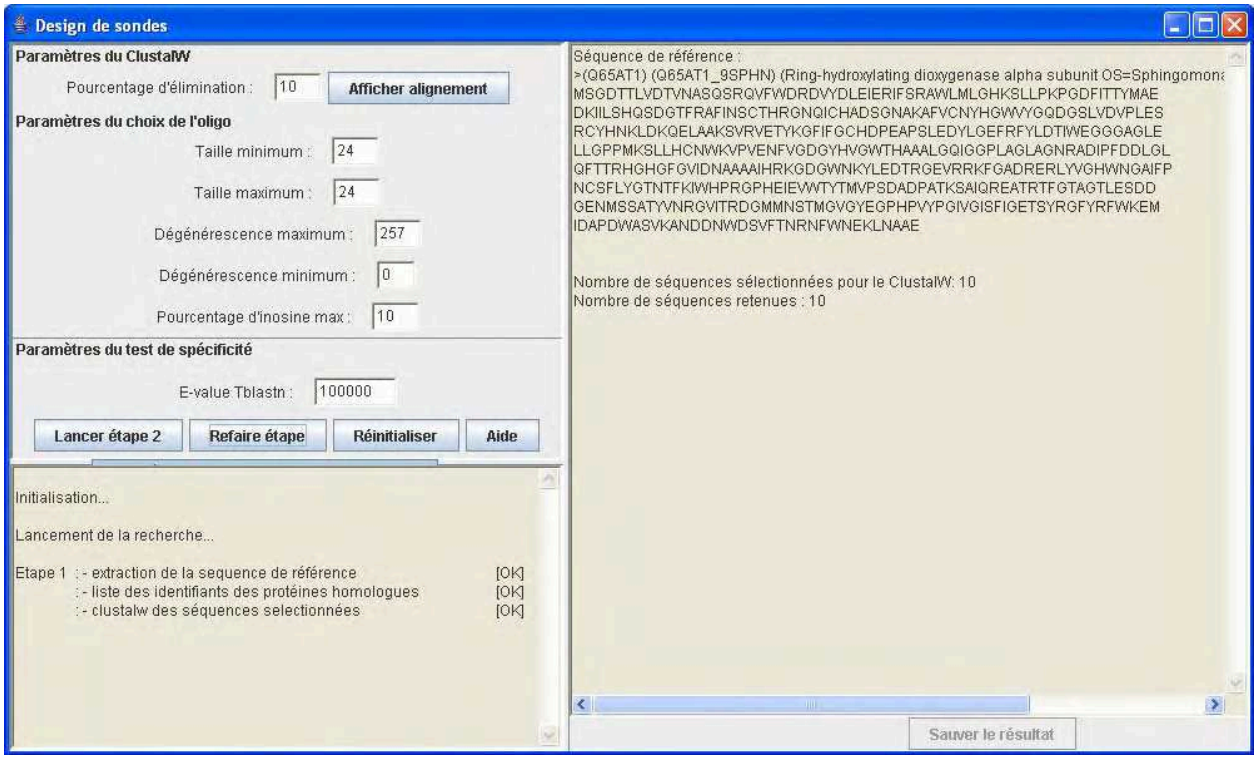


Figure 40 : Interface graphique de Metabolic Design pour la détermination de sondes exploratoires.

Paramètres de détermination des sondes : (1) Pourcentage d'élimination, pour l'étape de ClustalW (définissant l'écart de taille toléré des séquences des protéines sélectionnées, par rapport à la taille de la séquence de la protéine de référence). L'alignement multiple avec les séquences sélectionnées, peut alors être visualisé via le bouton : 'Afficher alignement' ; (2) taille des sondes, valeurs maximales de dégénérescence et du pourcentage de composés Inosine. (3) E-value pour l'évaluation de la spécificité des sondes. Trois boutons sont disponibles afin de lancer l'étape suivante (ici bouton 'Lancer étape 2'), de refaire l'étape en changeant les paramètres (bouton 'Refaire étape'), ou de réinitialiser le design (bouton 'Réinitialiser'). Finalement une aide sur le paramétrage et sur les différentes analyses est disponible via le bouton 'Aide'

La partie droite assure l'affichage des différentes informations pour l'utilisateur. Dans l'exemple, aucune des séquences des protéines sélectionnées n'a été éliminée pour la réalisation de l'alignement, car aucune n'était 10 % supérieures ou inférieures à la taille de la séquence de référence.

la détermination des sondes exploratoires, en utilisant les séquences de la protéine de référence et des protéines similaires (Figure 40).

2.5. Module de détermination de sondes exploratoires

Ce module, implémenté en JAVA pour l'interface graphique, et en PERL pour le traitement et la mise en forme des données, a été développée par Xavier Brotel durant son stage d'IUT Génie Biologique, option Bioinformatique, au sein de l'équipe de recherche.

La séquence de la protéine de référence, et les séquences protéiques sélectionnées par l'utilisateur, constitueront le point de départ de la détermination des sondes. Au lancement du module de design, l'utilisateur accède à une nouvelle fenêtre interne (Figure 40). A partir de là, les trois étapes qui suivent seront lancées indépendamment par l'utilisateur, et chaque étape pourra être effectuée plusieurs fois avec différents paramètres, grâce au bouton 'Refaire Etape'. La première étape consiste à effectuer un alignement multiple des séquences protéiques sélectionnées, la seconde étape permet le design des sondes vérifiant des caractéristiques définies par l'utilisateur (taille, pourcentage maximal d'Inosine, seuil maximal de dégénérescence), et la dernière étape permet d'apprécier la spécificité de chaque sonde par la comparaison avec la base de données Base de Données Spe.

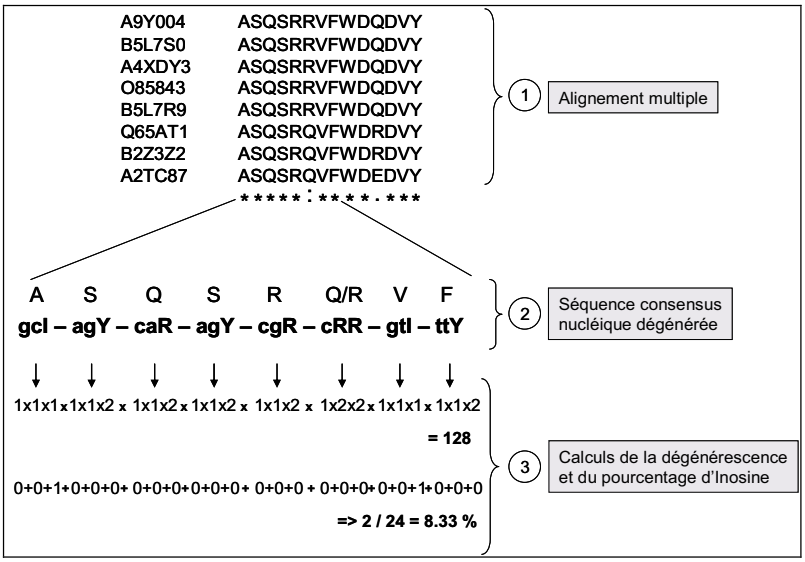
2.5.1. Alignement multiple des séquences protéiques sélectionnées

L'alignement des séquences protéiques sélectionnées est réalisé par le programme ClustalW. Cependant, pour obtenir un alignement de bonne qualité, un premier paramètre, défini par l'utilisateur, a été mis en place (nommé pourcentage d'élimination). En effet, pour réduire le nombre d'insertions et de délétions (ou indels) dans l'alignement multiple, il est nécessaire d'exclure les séquences dont la longueur est trop divergente par rapport à celle de la séquence de référence. Un message indique après ce test le nombre de séquences retenues, et effectivement alignées. Enfin, il est possible de visualiser cet alignement avec le bouton 'Afficher alignement'.

2.5.2. Détermination des sondes candidates

La sélection des sondes candidates peut se diviser en deux étapes : (1) la détermination de la séquence nucléotidique consensus dégénérée, basée sur l'alignement multiple protéique, et, (2) la sélection des sondes candidates répondant à différents paramètres définis par l'utilisateur (Annexe 2).

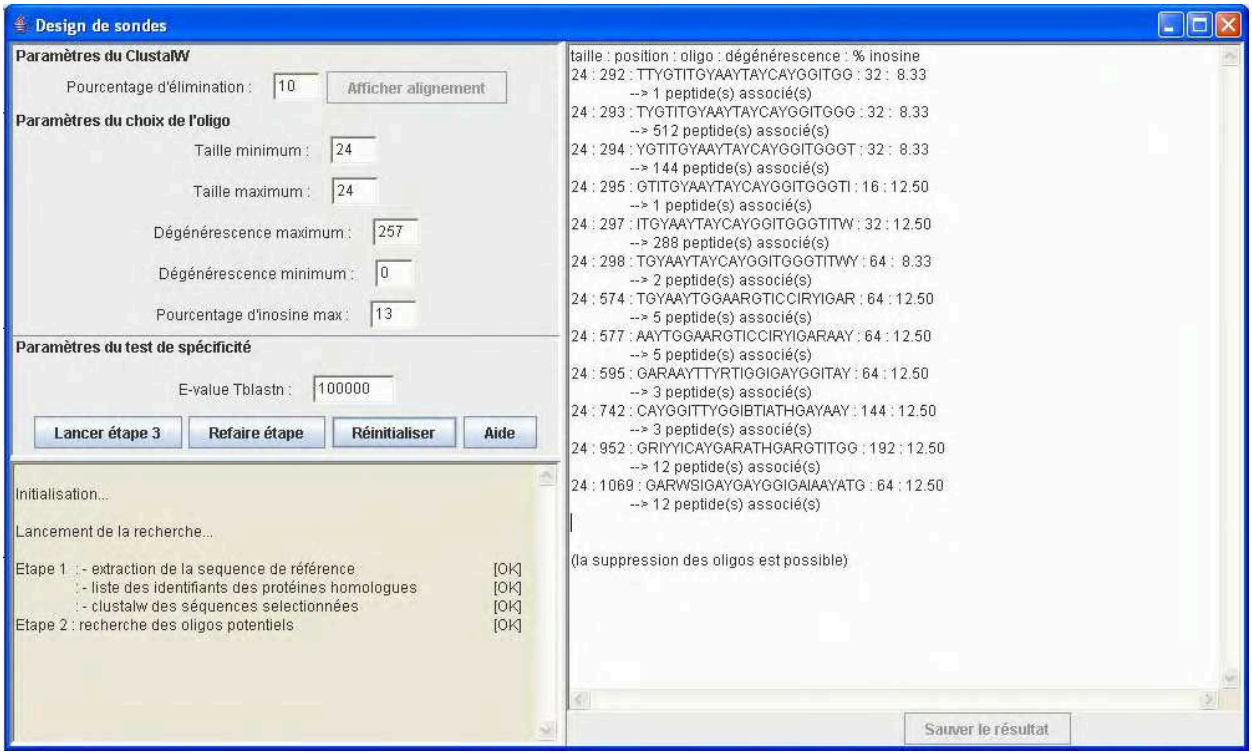
(1) La séquence nucléotidique consensus dégénérée est déterminée à partir de l'alignement multiple protéique obtenu par le logiciel ClustalW. Pour cela, une étape de traduction



<u>Figure 41 :</u> Stratégie de détermination des sondes dégénérées candidates à partir d'un alignement multiple protéique et de calcul de la dégénérescence et du taux d'Inosine.

<u>Nomenclature:</u> **R**: A et G; **Y**: C et T; **I**: A, C, G et T.

La dégénérescence traduit le nombre de sondes spécifiques pouvant être déduite de la sonde dégénérée, en excluant dans le calcul les résidus Inosines.



<u>Figure 42 :</u> Interface graphique de Metabolic Design présentant les résultats de sélection des sondes dégénérées candidates.

Pour chaque sonde, Metabolic Design donne la taille de la sonde, la position de cette sonde sur la séquence consensus, la séquence de la sonde, sa dégénérescence et sa composition en Inosine. Enfin, le nombre de séquences peptidiques codées pour chaque sonde dégénérée est indiqué, permettant de connaître le nombre d'itérations tBLASTn qui seront effectués pour l'estimation de la spécificité.

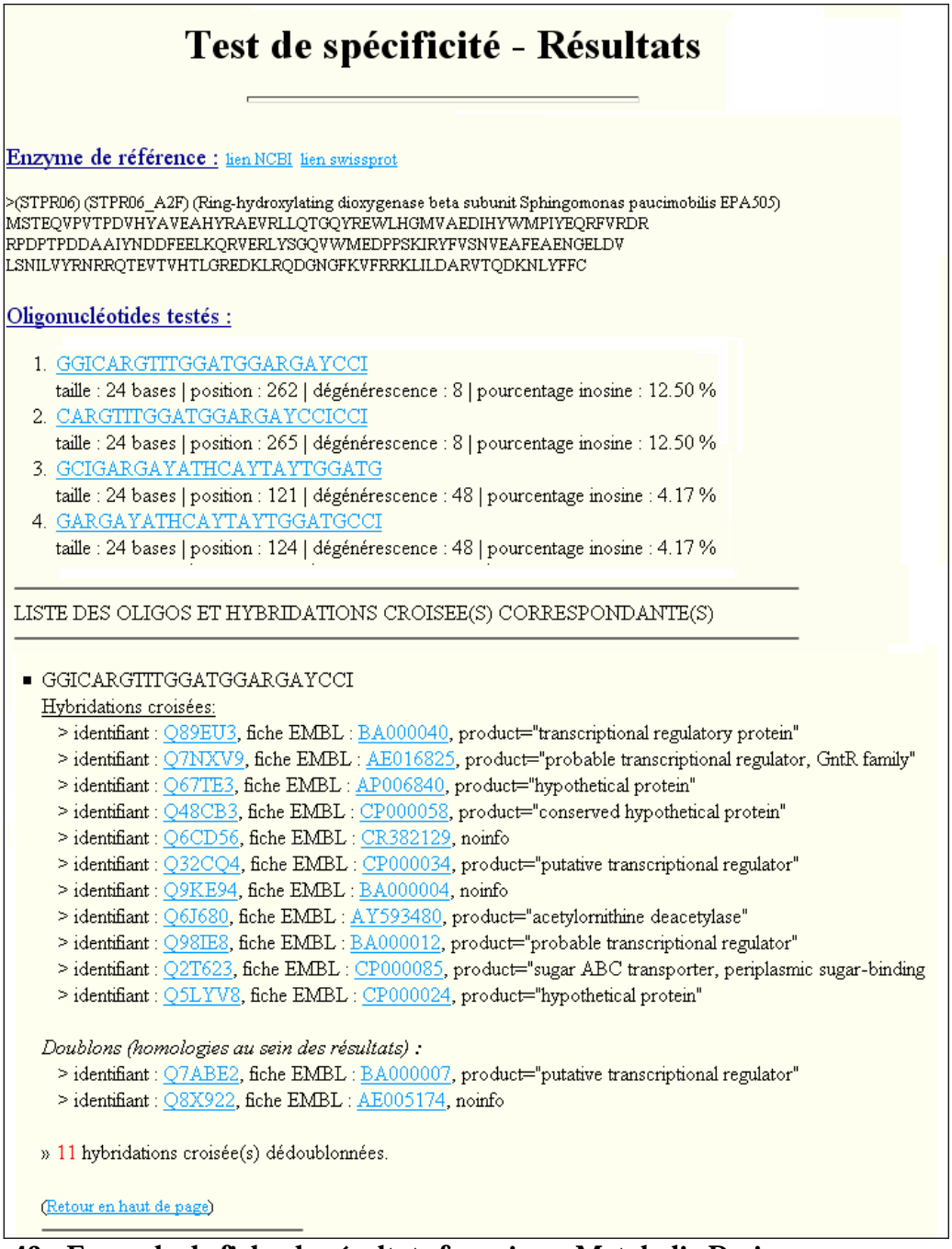
inverse est mise en place pour chaque site moléculaire de l'alignement protéique, en considérant la dégénérescence du code génétique, selon l'exemple donné dans la Figure 41. Cette traduction inverse est effectuée en s'appuyant sur le code IUPAC (nomenclature mise en place par « l'International Union of Pure and Applied Chemistry ») qui permet de formaliser l'écriture des séquences dégénérées.

(2) La sélection des sondes candidates se base sur plusieurs paramètres définis par l'utilisateur, et ajustables à façon. Il s'agit de la taille minimum et maximum des sondes, du taux de dégénérescence maximum accepté, et du pourcentage maximum d'Inosine entrant dans la composition de chaque sonde. La dégénérescence traduit le nombre de sondes spécifiques pouvant être déduite de la sonde dégénérée. Lorsque les paramètres ont été définis par l'utilisateur, la totalité de la séquence nucléique consensus dégénérée est analysée. En effet, chaque sonde dégénérée potentielle est évaluée, en se basant sur la taille spécifiée par l'utilisateur. Par exemple, la sonde de 18-mers de la position 1 à 18 sur la séquence consensus est analysée, puis celle de 2 à 19, etc.…, et cela sur toute la séquence consensus. Pour chaque oligonucléotide ainsi considéré, tous les paramètres sont évalués, et seuls ceux répondant aux critères définis sont conservés et listés. Toutes les sondes retenues sont ensuite affichées dans la fenêtre interne de résultat (Figure 42). Les paramètres calculés pour chaque sonde sont également indiqués avec d'autres informations (la position de la sonde sur la séquence consensus, et le nombre d'oligopeptides découlant de la séquence de la sonde dégénérée).

2.5.3. Estimation des hybridations croisées in silico

Cette étape permet de rechercher pour chaque sonde les hybridations croisées potentielles (Annexe 3). Notons que l'utilisateur peut supprimer les sondes qu'il ne souhaite pas analyser. La recherche est effectuée en comparant par BLAST les séquences des sondes dégénérées, avec les séquences de la base de données BdD Spe. Toutefois, afin de limiter le nombre d'itérations de BLAST réalisées, la comparaison est effectuée par un tBLASTn (le seuil d'E-value étant défini par l'utilisateur). En effet, via l'approche tBLASTn, tous les oligopeptides découlant de la séquence des sondes exploratoires sont définis. Les combinaisons de séquences protéiques potentiellement absentes dans les séquences protéiques utilisées pour faire l'alignement multiple vont donc également être générées.

Pour chaque séquence présentant une similarité de séquence avec l'un des oligopeptides déduit de la sonde dégénérée, plusieurs étapes d'analyse sont effectuées avant de considérer le résultat comme une hybridation croisée potentielle (Annexe 3) :



Test de spécificité - Résultats

Enzyme de référence : lien NCBI lien swissprot

>(STPR06) (STPR06_A2F) (Ring-hydroxylating dioxygenase beta subunit Sphingomonas paucimobilis EPA505)
MSTEQVPVTPDVHYAVEAHYRAEVRLLQTGQYREWLHGMVAEDIHYWMPIYEQRFVRDR
RPDPTPDDAAIYNDDFEELKQRVERLYSGQVWMEDPPSKIRYFVSNVEAFEAENGELDV
LSNILVYRNRRQTEVTVHTLGREDKLRQDGNGFKVFRRKLILDARVTQDKNLYFFC

Oligonucléotides testés :

1. GGICARGTITGGATGGARGAYCCI
   taille : 24 bases | position : 262 | dégénérescence : 8 | pourcentage inosine : 12.50 %
2. CARGTITGGATGGARGAYCCICCI
   taille : 24 bases | position : 265 | dégénérescence : 8 | pourcentage inosine : 12.50 %
3. GCIGARGAYATHCAYTAYTGGATG
   taille : 24 bases | position : 121 | dégénérescence : 48 | pourcentage inosine : 4.17 %
4. GARGAYATHCAYTAYTGGATGCCI
   taille : 24 bases | position : 124 | dégénérescence : 48 | pourcentage inosine : 4.17 %

LISTE DES OLIGOS ET HYBRIDATIONS CROISEE(S) CORRESPONDANTE(S)

- GGICARGTITGGATGGARGAYCCI

Hybridations croisées:
> identifiant : Q89EU3, fiche EMBL : BA000040, product="transcriptional regulatory protein"
> identifiant : Q7NXV9, fiche EMBL : AE016825, product="probable transcriptional regulator, GntR family"
> identifiant : Q67TE3, fiche EMBL : AP006840, product="hypothetical protein"
> identifiant : Q48CB3, fiche EMBL : CP000058, product="conserved hypothetical protein"
> identifiant : Q6CD56, fiche EMBL : CR382129, noinfo
> identifiant : Q32CQ4, fiche EMBL : CP000034, product="putative transcriptional regulator"
> identifiant : Q9KE94, fiche EMBL : BA000004, noinfo
> identifiant : Q6J680, fiche EMBL : AY593480, product="acetylornithine deacetylase"
> identifiant : Q98IE8, fiche EMBL : BA000012, product="probable transcriptional regulator"
> identifiant : Q2T623, fiche EMBL : CP000085, product="sugar ABC transporter, periplasmic sugar-binding
> identifiant : Q5LYV8, fiche EMBL : CP000024, product="hypothetical protein"

Doublons (homologies au sein des résultats) :
> identifiant : Q7ABE2, fiche EMBL : BA000007, product="putative transcriptional regulator"
> identifiant : Q8X922, fiche EMBL : AE005174, noinfo

» 11 hybridations croisée(s) dédoublonnées.

(Retour en haut de page)

Figure 43 : Exemple de fiche de résultats fourni par Metabolic Design.
La page HTML, automatiquement créée par le logiciel Metabolic Design donne toutes les informations sur les sondes testées et les hybridations croisées possibles. Elle reprend la séquence protéique de l'enzyme de référence, ainsi les liens de cette séquence sur NCBI et Swiss-Prot. Toutes les séquences engendrant potentiellement des hybridations croisées pour chaque sonde dégénérée sont regroupées en « familles », afin de supprimer les doublons potentiels (et de ne comptabiliser qu'une hybridation croisée potentielle par « famille »). La page HTML précise également ces séquences non comptabilisées, appelées 'Doublons'.

(1) Tout d'abord, Metabolic Design récupère le numéro d'accession de la séquence protéique présentant une hybridation croisée potentielle. Il vérifie alors si cet identifiant n'est pas dans la liste de séquences similaires extraites lors de la fouille de données initiale avec la protéine de référence.

(2) Il vérifie également si cet identifiant protéique ne correspond pas à une hybridation croisée déjà retrouvée lors d'autres itérations tBLASTn (cette vérification dynamique permet d'accélérer le temps de recherche des hybridations croisées).

(3) Si l'identifiant protéique n'est pas retrouvé dans ces deux listes, Metabolic Design récupère la séquence du gène codant la protéine engendrant potentiellement une hybridation croisée. Cette séquence nucléique est comparée à la séquence de la protéine de référence, par une étape de BLASTx. L'E-value obtenue détermine alors si la protéine est considérée comme une protéine de la même « famille », ou comme une séquence engendrant une hybridation croisée potentielle (le seuil d'E-value étant fixé à $1.e^{-10}$).

(4) Si la séquence n'est pas de la même famille, les critères de Kane, décrits précédemment, sont alors évalués, entre la séquence du gène correspondant, et la séquence de la sonde par une recherche de similarité BLASTn. Le résultat est alors sauvegardé, et une nouvelle hybridation croisée potentielle est analysée.

A la fin de l'analyse, toutes les séquences engendrant potentiellement des hybridations croisées pour chaque sonde dégénérée sont enfin regroupées en « familles », afin de supprimer les doublons potentiels (et de ne comptabiliser qu'une hybridation croisée potentielle par « famille »). Cette étape de regroupement est effectuée par une approche BLASTn. Le seuil d'E-value définissant une « famille » est fixé à $10.e^{-10}$. Chaque séquence est comparée aux autres par BLASTn, chaque « famille » formée est éliminée pour l'itération BLASTn suivante, jusqu'à ce qu'il ne reste plus de séquences à analyser. Les résultats sont stockés au fur et à mesure pour afficher les résultats.

En effet, suite à ces analyses, un fichier « résultat », au format HTML, est crée. Ce fichier liste toutes les sondes exploratoires testées ainsi que leurs caractéristiques (taille, dégénérescence, composition en Inosine, position sur la séquence consensus). Ce fichier liste également, pour chaque sonde exploratoire déterminée, les différentes séquences engendrant potentiellement des hybridations croisées. Enfin, des liens hypertextes sont disponibles, permettant d'afficher directement les fiches de chaque séquence nucléique, à partir des sites de l'EBI et d'UNIPROT, engendrant des hybridations croisées potentielles (Figure 43).

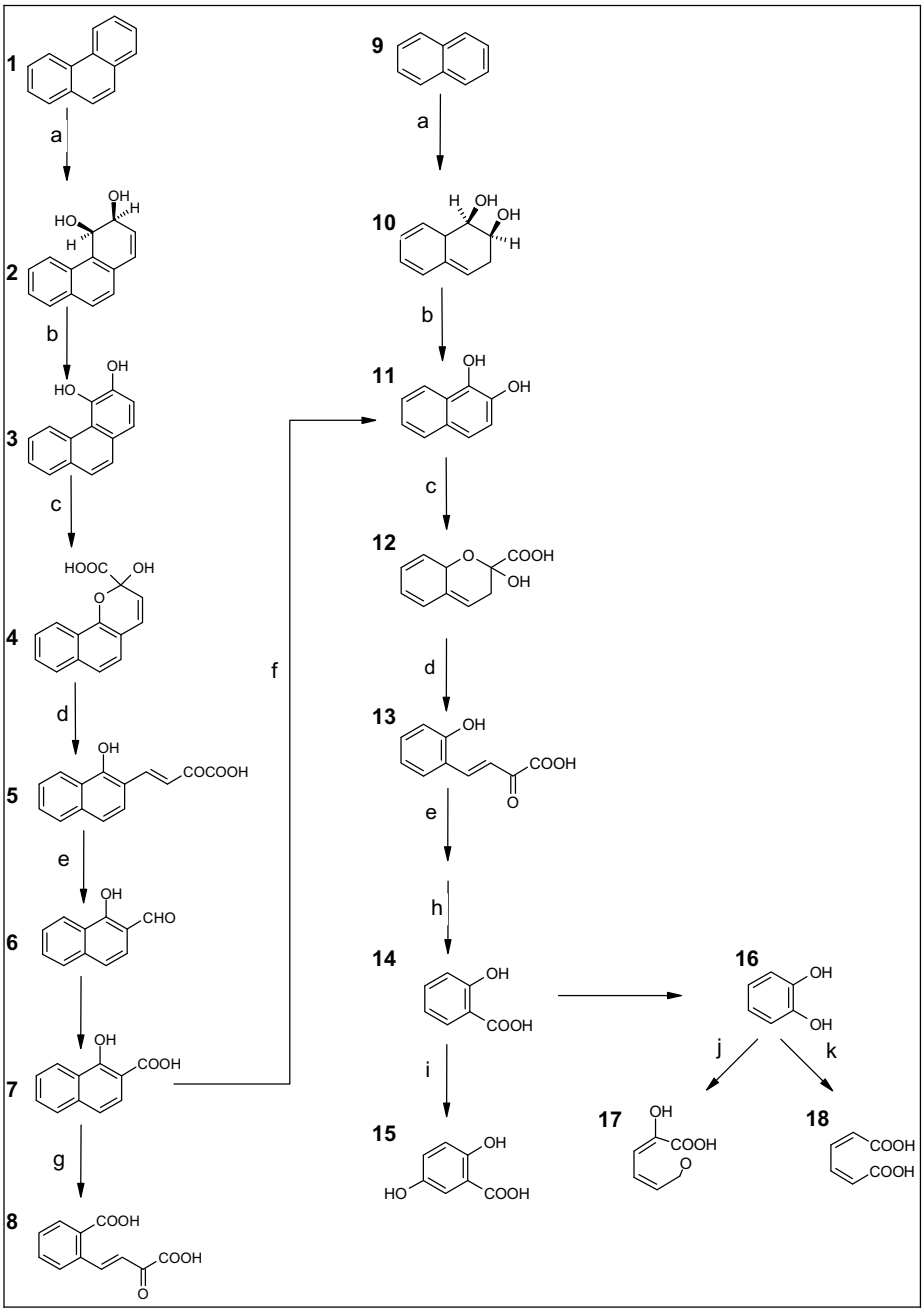


Figure 44 : Etapes enzymatiques des différentes voies de dégradation des HAP ciblées avec la biopuce fonctionnelle.

Les flèches sans lettres peuvent représenter une ou plusieurs étapes enzymatiques et ne sont pas ciblées.

Composés : (**1**) Phénanthrène ; (**2**) *cis*-3,4-phénanthrène dihydrodiol ; (**3**) 3,4-dihydroxyphénanthrène ; (**4**) 2-hydroxybenzo(*h*)chromène-2-carboxylate ; (**5**) 4-(1-hydroxynapht-2-yl)-2-oxobut-3-énoate ; (**6**) 1-hydroxy-2-naphtaldéhyde ; (**7**) 1-hydroxy-2-napthoate ; (**8**) *trans-o*-carboxybenzylidène pyruvate ; (**9**) naphtalène ; (**10**) *cis*-1,2-naphtalène dihydrodiol ; (**11**) 1,2-dihydroxynaphtalène ; (**12**) 2-hydroxychromène-2-carboxylate; (**13**) *trans-o*-hydroxybenzylidène pyruvate ; (**14**) salicylate ; (**15**) gentisate ; (**16**) catéchol ; (**17**) 2-hydroxymuconate semi aldéhyde ; (**18**) muconate.

Enzymes ciblées et gènes correspondants : (a) Dioxygénase initiale (*phnA1aA2a* correspondant respectivement aux sous-unités d'oxygénase α et β, *bphA3* à la ferrédoxine et *ahdA4* à la ferrédoxine réductase ; (b) dihydrodiol déshydrogénase (*bphB*) ; (c) dihydroxynaphtalène dioxygénase (*bphC*) ; (d) 2-hydroxychromène-2-carboxylate isomérase

3. Détermination et sélection des sondes pour la biopuce ADN métabolique ciblant les voies de dégradation des HAP

3.1. Voies métaboliques ciblées et gènes impliqués

Metabolic Design permet de définir des sondes pour biopuce ADN capables d'appréhender toute la diversité génique des familles de gènes ciblés, et donc les potentialités métaboliques d'un écosystème donné. L'objectif a été de définir des sondes ciblant les gènes codant des enzymes impliquées dans les voies métaboliques de dégradation de divers HAP, et d'autres molécules aromatiques. L'intérêt s'est donc porté sur les enzymes clés des voies considérées (Demaneche *et al.*, 2004; Ferraro *et al.*, 2005; Jakoncic *et al.*, 2007a, b; Kweon *et al.*, 2008; Stolz, 2009).

Ainsi, un total de 39 gènes impliqués ou potentiellement impliqués dans les mécanismes de dégradation de plusieurs HAP (notamment le phénanthrène, le fluoranthène ou le naphtalène), mais aussi des hydrocarbures aromatiques comme le biphényle, le benzoate ou les xylènes ont été ciblés via l'approche Metabolic Design, en se basant sur les informations disponibles pour les genres *Sphingomonas, Pseudomonas, Ralstonia* et *Nocardioides*.

3.1.1. Les voies métaboliques du phénanthrène

Les voies de minéralisation du phénanthrène ayant été fortement étudiées, différentes enzymes vont être ciblées (voir paragraphe *La dégradation bactérienne des HAP*, page 32. Ainsi, les enzymes de la voie commune (qui sont également impliquées dans les voies de dégradation d'autres HAP), ont tout d'abord été sélectionnées : (i) la dioxygénase responsable de l'attaque initiale du phénanthrène (composée de plusieurs sous-unités), (ii) la déshydrogénase, (iii) la dioxygénase clivant en position *meta* le 3,4-dihydroxyphénanthrène, (iv) l'isomérase et (v) l'hydratase-aldolase (Figure 44 ci-contre et Tableau 15 page suivante).

Ces enzymes interviennent dans les premières étapes de la dégradation du phénanthrène, mais aussi du naphtalène (Figure 44) Ces étapes enzymatiques conduisent à la formation de l'acide-1-hydroxy-2-naphtoïque en présence de phénanthrène, point de départ de deux voies de dégradation distinctes : celle de Evans et celle de Kiyohara (Kiyohara et Nagao, 1978; Iwabuchi et Harayama, 1998b; Cho *et al.*, 2005). Les enzymes caractéristiques de chacune de ces voies ont donc également été ciblées (Tableau 15). Pour la voie de Kiyohara, il s'agit de l'enzyme 1-hydroxy-2-naphtoate dioxygénase (ou PhdI) dont le gène a été isolé initialement chez *Nocardioides* sp. KP7 (voir Tableau 15) (Iwabuchi et Harayama, 1998b). En

(*nahD*) ; (e) dihydroxybenzylpyruvate aldolase (*nahE*) ; (f) salicylate oxygénase (*ahdA1cA2c* correspondant respectivement aux sous-unités α et β d'oxygénase, *bphA3* à la ferrédoxine et *ahdA4* à la ferrédoxine réductase) ; (g) 1-hydroxy-2-napthoate dioxygénase (*phdI*) ; (h) salicylaldéhyde déshydrogénase NAD-dépendante (*nahF*) ; (i) salicylate-5-hydroxylase (*nagG* et *nagH*) ainsi que le régulateur de cette voie (codé par *nahR*), (j) catéchol 2,3-dioxygénase (*xylE*) ; (k) catéchol 1,2-dioxygénase (*catA*), ainsi que le régulateur potentiel de cette voie (*catR*).

Tableau 15 : Gènes ciblés avec l'approche Metabolic Design pour l'étude des voies de dégradation du phénanthrène et du naphtalène, et des voies « basses » de dégradation.

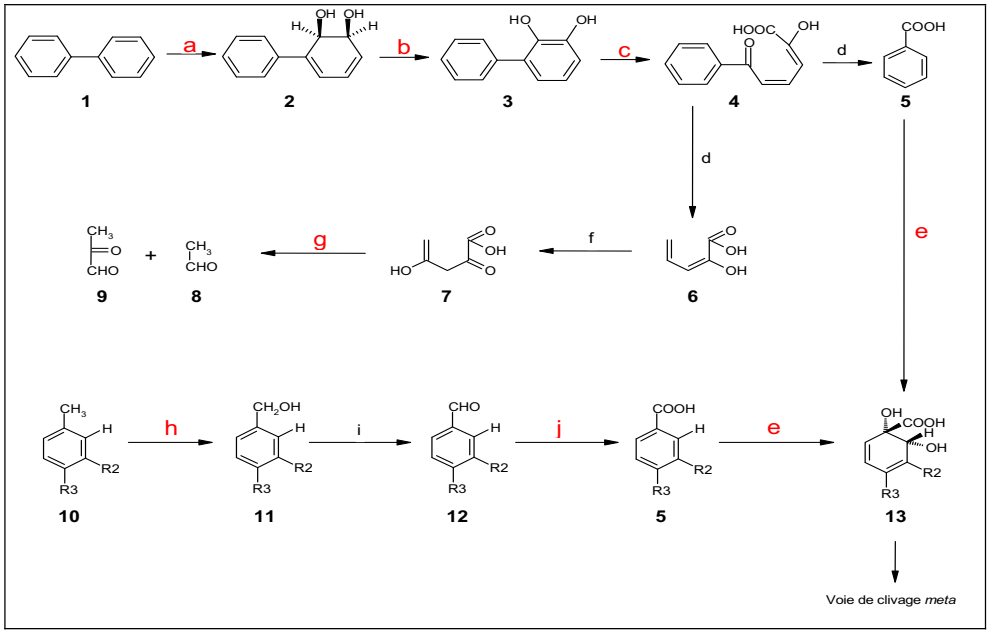
Gène	Enzyme	Organisme	Numéro d'accession	Voie métabolique	Référence(s)
phnA1a	Sous-unité α de la dioxygénase initiale	*Sphingomonas* sp. CHY-1	Q65AT1	Phénanthrène Naphtalène	(Demaneche *et al.*, 2004)
phnA2a	Sous-unité β de la dioxygénase initiale	*Sphingomonas* sp. CHY-1	Q65AT0	Phénanthrène Naphtalène	(Demaneche *et al.*, 2004)
bphA3	Sous-unité de ferrédoxine	*Sphingomonas yanoikuyae*	A2TC31	Phénanthrène Naphtalène	(Cho *et al.*, 2005; Ní Chadhain *et al.*, 2007)
ahdA4	Ferrédoxine réductase	*Sphingomonas yanoikuyae*	A2TC59	Phénanthrène Naphtalène	(Cho *et al.*, 2005; Ní Chadhain *et al.*, 2007)
bphB	*cis*-dihydrodiol déshydrogénase	*Sphingobium xenophagum*	Q9X9Q9	Phénanthrène Naphtalène	(Keck *et al.*, 2006)
bphC	Dihydroxynaphtalène dioxygénase	*Sphingobium xenophagum*	P74836	Phénanthrène Naphtalène	(Keck *et al.*, 2006)
nahD	2-hydroxychromène-2-carboxylate isomérase	*Sphingomonas* sp. P2	Q83VL3	Phénanthrène Naphtalène	(Kim *et al.*, 1997; Cho *et al.*, 2005)
nahE	Dihydroxybenzylpyruvate aldolase	*Sphingomonas aromaticivorans*	O85960	Phénanthrène Naphtalène	(Romine *et al.*, 1999b)
phdI	1-hydroxy-2-naphthoate dioxygénase	*Nocardioides* sp. KP7	Q9FBF3	Kiyohara	(Iwabuchi et Harayama, 1998b)
ahdA1c	Grande sous-unité d'oxygénase	*Sphingomonas* sp. P2	Q83VL2	Evans	(Pinyakong *et al.*, 2003b; Cho *et al.*, 2005)
ahdA2c	Petite sous-unité d'oxygénase	*Sphingomonas* sp. P2	Q83VL1	Evans	(Pinyakong *et al.*, 2003b; Cho *et al.*, 2005)
nahF	Salicylaldéhyde déshydrogénase	*Sphingomonas aromaticivorans*	O86001	Naphtalène	(Romine *et al.*, 1999b; Stolz, 2009)
xylE	catéchol 2,3-dioxygénase	*Sphingomonas aromaticivorans*	O85982	Clivage *meta*	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
catA	catéchol 1,2-dioxygénase	*Sphingomonas* sp. KP1	Q0KJT4	Clivage *ortho*	(Rothmel *et al.*, 1990; McFall *et al.*, 1998)
catR	Régulateur opéron *cat*	*Sphingomonas* sp. KP1	Q0KJT7	Clivage *ortho* Régulateur	(Rothmel *et al.*, 1990; McFall *et al.*, 1998)
nagG	Grande sous-unité du salicylate 5-hydroxylase	*Ralstonia* sp. U2	O52379	Gentisate	(Zhou *et al.*, 2001)
nagH	Petite sous-unité du salicylate 5-hydroxylase	*Ralstonia* sp. U2	O52380	Gentisate	(Zhou *et al.*, 2001)
nagR	Régulateur opéron *nag*	*Ralstonia* sp. U2	Q9EXL7	Gentisate Régulateur	(Jones *et al.*, 2003)

ce qui concerne la voie de Evans, l'enzyme multimérique transformant l'acide 1-hydroxy-2-naphthoïque, en 1,2-dihydroxynaphtalène, a été choisie, car permettant de rejoindre la voie de dégradation du naphtalène (Figure 44). Ce complexe multimérique est codé chez *Sphingomonas yanoikuyae* B1 et *Sphingomonas* sp. P2 par les gènes *ahdA2cA1c, bphA3* et *ahdA4* (voir Tableau 15) (Pinyakong *et al.*, 2003b; Cho *et al.*, 2005). L'action de cette enzyme permettant donc de rejoindre la voie de dégradation du naphtalène, il a été également nécessaire de cibler sa voie de dégradation de manière spécifique. L'enzyme NahF (une salicylaldéhyde déshydrogénase), a été choisie, se situant à la fin de la voie de dégradation du naphtalène (Figure 44), et existant chez de nombreux organismes, notamment les *Sphingomonas* et les *Pseudomonas*, proche de notre souche modèle (Vandecasteele, 2005; Stolz, 2009)

Enfin, le salicylate potentiellement formé par ces voies de dégradation, peut alors rejoindre trois voies de minéralisation différentes : les voies de clivage dite *meta* et *ortho* (en formant du catéchol) et la voie de dégradation du gentisate (Figure 44). Des enzymes spécifiques de chacune de ces trois voies ont donc été ciblées, de manière à pouvoir connaître la voie de minéralisation du salicylate mise en jeu (Tableau 15). Il a été décidé de cibler les enzymes les plus connues des voies de clivage *meta* et *ortho* (la catéchol 2,3-dioxygénase, ou XylE, et la catéchol 1,2-dioxygénase, ou CatA, enzymes spécifiques des voies de clivage *meta* et *ortho*), mais aussi le régulateur potentiel de la voie de clivage *ortho*, nommé CatR (Rothmel *et al.*, 1990; McFall *et al.*, 1998). En ce qui concerne la voie du gentisate, l'enzyme multimérique appelée salicylate-5-hydroxylase, caractérisée chez *Ralstonia* sp. U2, a été choisie (Tableau 15). Les deux sous-unités d'oxygénase de cette dernière ont donc été utilisées : NagG et NagH (Zhou *et al.*, 2001). La protéine NagR, régulant l'expression des gènes codant les protéines impliquées dans la suite de la voie du gentisate chez *Ralstonia* sp. U2 a également été ciblée (Jones *et al.*, 2003).

3.1.2. Les voies métaboliques d'autres composés aromatiques

Chez certaines espèces appartenant au genre *Sphingomonas*, les gènes codant des enzymes impliquées dans la dégradation de composés monoaromatiques (comme le toluène, le benzoate ou les xylènes) sont organisés au sein de mêmes opérons que les gènes codant des enzymes impliquées dans la dégradation des HAP (Romine *et al.*, 1999b). De plus, la voie de dégradation du biphényle, un composé aromatique également dégradé par les microorganismes a été ciblée, pour compléter la biopuce.



<u>Figure 45 :</u> Voies de dégradation du biphényle, du toluène et des xylènes.
Composés : (**1**) Biphényle ; (**2**) *cis*-2,3-biphényle dihydrodiol ; (**3**) 2,3-dihydroxybiphényle ; (**4**) 2-Hydroxy-6-oxo-6-phénylhexa-2,4-diénoate ; (**5**) benzoate ; (**6**) *cis*-2-Hydroxypenta-2,4-diénoate ; (**7**) 4-Hydroxy-2-oxovalérate ; (**8**) acétaldéhyde ; (**9**) pyruvate ; (**10**) Toluène ; (**11**); alcool benzylique ; (**12**) benzaldéhyde ; (**13**) benzoate dihydrodiol. Les groupements R2 et R3 peuvent différer selon si l'on considère du toluène, du *o*-xylène, du *m*-xylène ou du p-xylène en substrat initial. Les étapes enzymatiques sont les mêmes pour ces différents composés.
Enzymes et nomenclature des gènes correspondants donnée chez *Sphingomonas yanoikuyae* B1 (en rouge apparaissent les gènes ciblés) : (a) Naphtalène-1,2-dioxygénase (*phnA1aA2A, bphA3, ahdA4*) ; (b) *cis*-1,2-naphtalène dihydrodiol déshydrogénase (*bphB*) ; (c) 1,2-dihydroxynaphtalène dioxygénase (*bphC*) ; (d) 2-Hydroxy-6-oxo-6-phénylhexa-2,4-diénoate hydrolase (*bphD*) ; (e) toluène dioxygénase (*xylXY, bphA3, ahdA4*) ; (f) 2-oxopent-4-énoate hydratase (*bphE*) ; (g) 4-hydroxy-2-oxovalérate aldolase (*bphF*) ;(h) xylène mono-oxygénase (*xylAM*) ; (i) alcool benzylique déshydrogénase (*xylB*) ; (j) benzaldéhyde déshydrogénase (*xylC*) (adaptée de Pinyakong *et al.*, 2003).

<u>Tableau 16 :</u> Gènes ciblés avec l'approche Metabolic Design pour les voies de dégradation d'autres composés aromatiques.
Pour chaque gène est précisé son nom, le nom de l'enzyme codée par ce gène, le nom de l'organisme source, le numéro d'accession de la séquence protéique utilisée pour le design, le substrat mis en jeu ainsi que l'enzyme ciblée, et les références bilbliographiques.

Gène	Enzyme	Organisme	Numéro d'accession	Substrat	Référence
xylM	Sous-unité de la monooxygénase du xylène	*Sphingomonas aromaticivorans*	O85970	Xylène Toluène	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
xylA	Sous-unité de la monooxygénase du xylène	*Sphingomonas aromaticivorans*	O85971	Xylène Toluène	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
xylC	Benzaldéhyde deshydrogénase	*Sphingomonas aromaticivorans*	O85973	Xylène Toluène	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
xylX	Sous-unité α de la toluène dioxygénase	*Sphingomonas yanoikuyae*	A2TC33	Benzoate Biphényle	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
xylY	Sous-unité β de la toluène dioxygénase	*Sphingomonas yanoikuyae*	A2TC34	Benzoate Biphényle	(Kim et Zylstra, 1999; Romine *et al.*, 1999b)
bphF	4-hydroxy-2-oxovalérate aldolase	*Pseudomonas* sp. KKS102	P51014	Biphényle	(Kikuchi *et al.*, 1994; Stolz, 2009)

Les enzymes les plus connues de ces voies ont donc été sélectionnées (Tableau 16 et Figure 45) (Vandecasteele, 2005; Stolz, 2009). Les enzymes sont : XylM et XylA (respectivement deux sous-unités de la mono-oxygénase du xylène), XylC (une benzaldéhyde déshydrogénase impliquée dans la dégradation du xylène et du toluène), XylX et XylY (les deux sous-unités de la toluène dioxygénase mise en jeu dans la dégradation du toluène et du benzoate), et enfin BphF (une 4-hydroxy-2-oxo-valérate aldolase, spécifique de la dégradation du biphényle, et permettant la formation de pyruvate et d'acétaldéhyde) (Tableau 16 et Figure 45).

3.1.3. Autres gènes ciblés

D'autres protéines ont également été sélectionnées, en raison de leur rôle potentiel dans la dégradation de composés aromatiques. Chez le genre *Sphingomonas*, les gènes codant des enzymes impliquées dans la dégradation des HAP, du biphényle et des hydrocarbures monoaromatiques sont le plus souvent localisés sur un plasmide (Pinyakong *et al.*, 2000; Pinyakong *et al.*, 2003a; Stolz, 2009). Or, il a également été montré que sur ce plasmide, de nombreuses sous-unités d'oxygénase étaient présentes (Stolz, 2009). Ces différentes enzymes confèrent probablement la capacité de pouvoir dégrader une large gamme de substrats aromatiques aux espèces du genre *Sphingomonas* (Stolz, 2009). Les spécificités de chacune de ces protéines n'ont cependant pas encore été décrites dans la littérature. Il a donc été décidé de cibler les six couples de petite et grande sous-unités d'oxygénase référencés au sein du plasmide de la souche *Sphingomonas aromaticivorans* F199, sur lequel on retrouve notamment les gènes spécifiques de la dégradation des HAP (Romine *et al.*, 1999b; Basta *et al.*, 2004) (Tableau 17 page suivante).

De même, au sein du plasmide de la souche *Sphingomonas aromaticivorans* sont présents plusieurs gènes codant des régulateurs potentiels. Six régulateurs, parmi ceux détectés, font partie de familles ou de sous-familles connues comme impliquées dans la régulation de la dégradation de molécules aromatiques (Romine *et al.*, 1999a). C'est pourquoi, tous les gènes codant pour ces régulateurs ont également été ciblés (Tableau 18 page suivante), de manière à déterminer si l'expression des gènes codant ces régulateurs est modulée par la présence ou l'absence de HAP.

Finalement, deux autres gènes ont été choisis, le gène *gyrB* (codant la sous-unité β de l'ADN gyrase, connue chez *Sphingomonas* sp. SKA58) et le gène *bphK* (codant une glutathione-S-transférase, connue chez *Sphingomonas* sp. P2). Le premier est un gène dit de ménage, dont l'expression est censée être constitutive. Le second gène a été démontré comme

<u>Tableau 17 :</u> Gènes ciblés avec l'approche Metabolic Design codant différentes (di-)oxygénases.
Toutes les séquences utilisées sont basées sur l'étude de Romine *et al.,* 1999b, sauf *bphA1d* et *bphA2d*, provenant de l'étude de Cho *et al.*, 2005.

Gène	Enzyme	Organisme	Numéro d'accession
bphA1a	Grande sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85964
bphA2a	Petite sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85965
bphA1b	Grande sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85966
bphA2b	Petite sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85967
bphA1d	Grande sous-unité d'oxygénase	*Sphingomonas* sp. P2	O85986
bphA2d	Petite sous-unité d'oxygénase	*Sphingomonas* sp. P2	O85985
bphA1e	Grande sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85959
bphA2e	Petite sous-unité d'oxygénase	*Sphingomonas aromaticivorans*	O85958

<u>Tableau 18 :</u> Gènes ciblés avec l'approche Metabolic Design codant différents régulateurs transcriptionnels présents sur le plasmide séquencé de la souche *Sphingomonas aromaticivorans* **F199.**
Toutes les séquences utilisées proviennent de la souche *Sphingomonas aromaticivorans* et sont basées sur l'étude de Romine *et al.,* 1999b.

Gène	Type de régulateur	Numéro d'accession
bphR	Régulateur associé avec Sigma-54 pour l'ARN polymérase	O85963
Orf007	Famille IclR	O85831
Orf158	Famille MarR	O85850
Orf569	Famille Sigma-24 pour l'ARN polymérase	O85900
Orf597	Famille MucR	O85904
Orf758	Famille GntR	O85926

exprimé en présence de HAP (Lloyd-Jones et Lau, 1997). Ces deux gènes serviront donc de contrôles internes. Ainsi, un total de 40 protéines a été ciblé (Annexe 4).

3.2. Stratégies de détermination de sondes appliquées

Afin de maximiser le nombre de sondes définies par Metabolic Design, deux stratégies différentes ont été utilisées. Ces deux stratégies ont été appliquées à partir du même alignement multiple, pour chacune des protéines étudiées. Dans les deux cas, la taille des sondes a été fixée à 24-mers, cette longueur représentant le meilleur compromis, entre la spécificité, et la sensibilité des sondes (Rimour *et al.*, 2005).

La première stratégie employée a consisté à autoriser une dégénérescence totale importante des sondes, notamment en permettant la présence d'un fort pourcentage d'Inosine (dégénérescence maximum sans tenir compte des Inosines : 129, pourcentage maximum d'Inosine : 25 %). Cependant, de nombreuses sondes montraient une trop forte dégénérescence totale (en considérant les Inosines), engendrant un trop grand nombre de sondes spécifiques à synthétiser sur les formats de biopuces *in situ* choisis. Aussi, une seconde stratégie a été mise en place, limitant la composition en Inosine (pourcentage maximum d'Inosine : 9 %), mais autorisant un seuil maximum de dégénérescence ne tenant pas compte des Inosines plus élevé (de 258). En utilisant ces nouveaux paramètres, un deuxième groupe de sondes a été défini pour chacune des 40 enzymes ciblées. La spécificité de ces deux jeux de sondes définies a ensuite été estimée.

Parmi ces deux jeux de sondes disponibles, les sondes qui seront utilisées ont été sélectionnées selon trois critères. (i) Le nombre d'hybridations croisées potentielles *in silico* détectées, et leur nature. Les sondes présentant le plus faible nombre d'hybridations croisées ont été conservées, mais le type de gène entraînant des hybridations croisées potentielles a aussi été considéré. Par exemple, si la majorité des hybridations croisées est engendrée par des gènes retrouvés uniquement au sein d'environnements anaérobies, il est possible d'émettre l'hypothèse qu'il y ait peu de chances de retrouver ces gènes dans les environnements aérobies étudiés. (ii) La valeur de la dégénérescence totale (en incluant les Inosines) de chaque sonde. (iii) Le nombre de régions géniques ciblées. Lorsque cela était possible, trois zones différentes ont été considérées, mais la plupart du temps, seules deux zones ont été conservées, et parfois une seule région. Cela a permis de sélectionner un total de 72 sondes dégénérées pour 40 gènes ciblés (Annexe 5). Enfin, toujours de manière à réduire la dégénérescence totale des sondes spécifiques, la dernière base de chaque sonde a été éliminée. En effet, la dernière base de chaque sonde est la troisième base d'un codon, et celle-ci présente

la plus forte dégénérescence au sein du code génétique. Au final la biopuce ADN fonctionnelle comporte 39 216 sondes non dégénérées.

3.3. Bilan

Grâce au module de visualisation de Metabolic Design, une reconstruction graphique des voies métaboliques étudiées a été réalisée. A partir de cette visualisation, une fouille de données a été effectuée, et les différentes séquences protéiques extraites ont permis de définir des sondes ciblant les gènes codant des enzymes ou des protéines régulatrices impliquées dans les voies métaboliques de dégradation de plusieurs composés aromatiques (dont principalement des HAP). Ainsi, 72 sondes exploratoires, pour les 40 enzymes ciblées des voies de dégradation ont été définies. Ces 72 sondes dégénérées représentent un total de 39 216 sondes spécifiques. L'avantage de cette approche est que ces sondes exploratoires permettent d'appréhender toute la diversité génique. En effet, ces dernières peuvent s'hybrider avec des séquences nucléiques connues, mais également avec des séquences encore non référencées au sein des bases de données mais codant la portion peptidique ciblée et donc potentiellement spécifique de la protéine étudiée. La validation de cette biopuce ADN a alors été entreprise.

Chapitre II : Validation de la biopuce métabolique exploratoire

1. Introduction

Afin de valider l'approche développée pour la détermination de sondes exploratoires sans *a priori* de séquences, l'utilisation d'une souche modèle connue pour dégrader divers HAP a été envisagée.

La souche sélectionnée, appelée *Sphingomonas paucimobilis* sp. EPA505, utilise le phénanthrène et le fluoranthène, deux HAP, comme source de carbone et d'énergie (Mueller *et al.*, 1990). Cependant, chez cette souche bactérienne, les enzymes impliquées dans les étapes de dégradation de ces composés n'ont pas été caractérisées. Seuls des fragments de gène codant une sous-unité de ferrédoxine (nommée *pbhB* dans l'étude, correspondant à *bphA3* dans notre nomenclature), une *trans*-*o*-hydroxybenzylidène pyruvate hydratase-aldolase (nommée *pbhC* dans l'étude, et correspondant à *nahE*) et une pyruvate phosphate dikinase ont été isolés (Story *et al.*, 2000). De plus, la séquence du gène codant une dihydroxybiphényle-2,3-diol 1,2-dioxygénase (nommée *pbhA*, et correspondant à *bphC*) a été complètement isolée durant cette même étude (Story *et al.*, 2000). Les trois séquences *pbhA, pbhB* et *pbhC* codent donc potentiellement pour des enzymes ciblées par les sondes sélectionnées avec Metabolic Design. Aucune autre donnée n'est disponible concernant les autres gènes codant pour des enzymes de dégradation des HAP. Dans ce contexte, le suivi de l'expression des gènes ciblés par la biopuce a été entrepris dans le cadre d'études de croissance de *Sphingomonas paucimobilis* sp. EPA505 avec comme seules sources de carbone et d'énergie différents HAP.

2. Suivi de croissance et de dégradation des HAP

2.1. Conditions de culture

Les conditions de culture de la souche *Sphingomonas paucimobilis* sp. EPA505 en présence de différents HAP, comme seule source de carbone et d'énergie ont été optimisées. Les arrêts d'oxygénation ayant un effet négatif sur la croissance bactérienne, un Erlenmeyer est ensemencé pour chaque prélèvement. Pour limiter les risques de contamination lors de

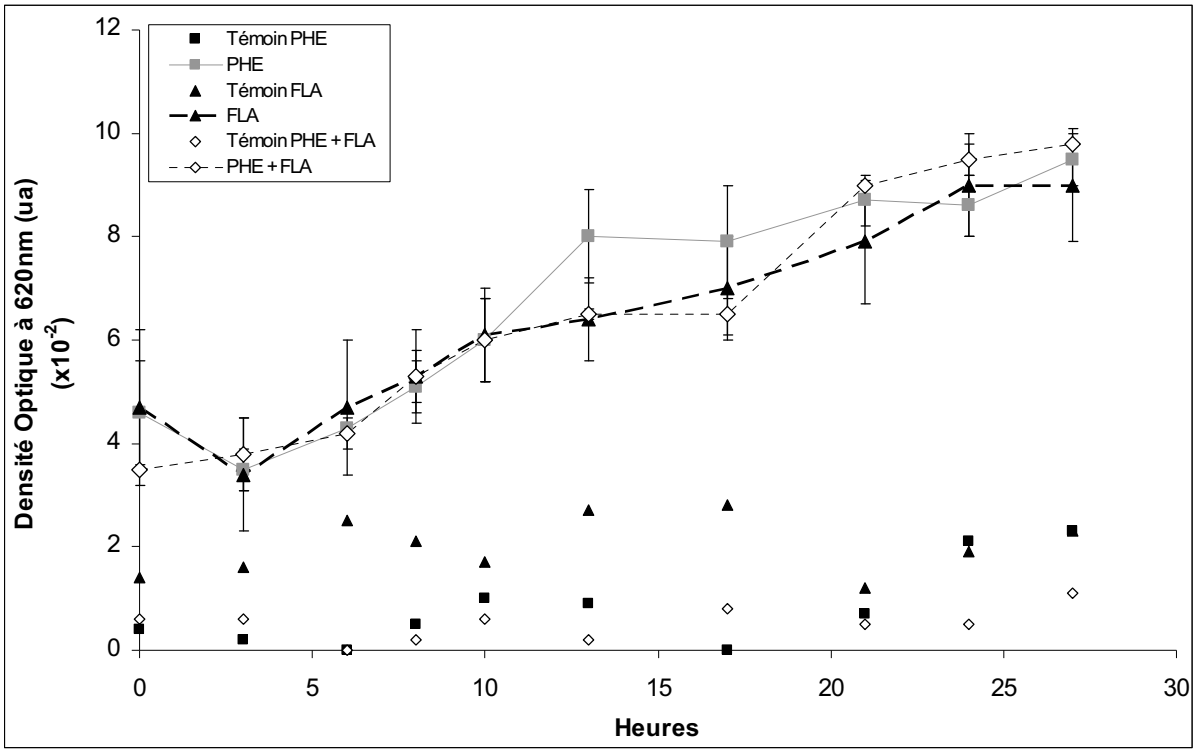


Figure 46 : Suivis de croissance réalisés par mesure d'absorbance à 620nm de la souche *S. paucimobilis* sp. EPA505 en présence de HAP comme seule source carbonée.
PHE : phénanthrène, FLA : fluoranthène. Les témoins ne contiennent aucun *inoculum* bactérien. Les barres d'erreur représentent les variations potentielles engendrées par l'hétérogénéité du milieu.

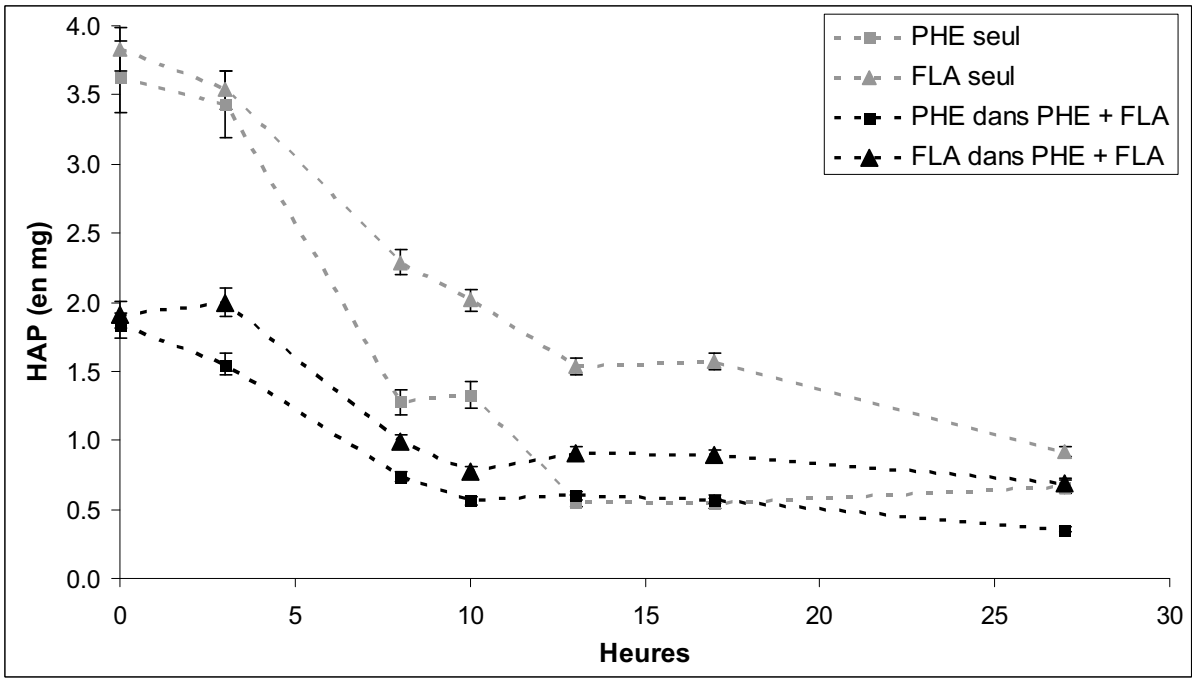


Figure 47 : Suivis de dégradation des HAP par CLHP, durant la croissance de la souche *Sphingomonas paucimobilis* sp. EPA505.
PHE : phénanthrène, FLA : fluoranthène. Les barres d'erreur représentent les pertes potentielles estimées durant les étapes d'extraction.

l'ajout des composés, de la streptomycine est également introduite à une concentration finale de 100µg/mL (Vanbroekhoven *et al.*, 2004).

2.2. Suivis des cinétiques de croissance bactérienne et de biodégradation des HAP.

Les capacités de croissance de la souche modèle EPA505 en présence de phénanthrène, de fluoranthène et d'un mélange de ces deux composés (50/50) comme seule source de carbone et d'énergie ont été suivies par des mesures d'absorbance à 620nm. Parallèlement, un suivi de la dégradation des HAP a été réalisé par des analyses CLHP.

En absence de HAP, les mesures de densité optique ne montrent aucune croissance, confirmant l'impossibilité de la souche EPA505 à utiliser le Tween 80 (un surfactant, ajouté pour une meilleure solubilisation des HAP en milieu aqueux), comme source de carbone et d'énergie. Il est également important de noter que les variations de densités optiques observées pour les différentes mesures des témoins négatifs (sans *inoculum* bactérien) résultent de l'hétérogénéité liée à la faible dissolution des HAP en milieu aqueux (Figure 46).

Les résultats obtenus montrent que la souche EPA505 est capable de croître en présence de phénanthrène ou de fluoranthène (Figure 46). Pour ces deux composés, les courbes de croissance révèlent un profil similaire. En effet, après une phase de latence de 3 heures, une phase de croissance est visible entre 3 et 13 heures pour le phénanthrène, et entre 3 et 17 heures pour le fluoranthène. Le suivi de dégradation des HAP montre que la disparition du phénanthrène a principalement lieu entre 3 et 8 heures de culture (Figure 47). En effet, 60 % du phénanthrène initialement présent est consommé après 8 heures. Cette dégradation se poursuit pour atteindre 84 % après 27 heures de culture. Pour le fluoranthène, 48 % est consommé après 8 heures et 75,9 % en fin de cinétique.

La souche EPA505 est également capable de croître en présence d'un mélange de phénanthrène et de fluoranthène (50/50). La croissance s'effectue en deux phases successives : la première entre 3 et 13 heures de culture, et la seconde entre 17 et 24 heures (Figure 46). Dans les deux cas, elles sont précédées par une phase de latence. Le suivi des sources carbonées par CLHP révèle qu'après 3 heures de culture, 15,5 % du phénanthrène est dégradé, alors que la quantité de fluoranthène reste stable (Figure 47). Puis, entre 3 et 17 heures de culture, une dégradation des deux composés est observée, avec 72,5 % de phénanthrène et 61 % de fluoranthène respectivement dégradés après 17 heures de culture. A la fin de la cinétique, ces valeurs atteignent 80,9 % pour le phénanthrène et 65,6 % pour le fluoranthène.

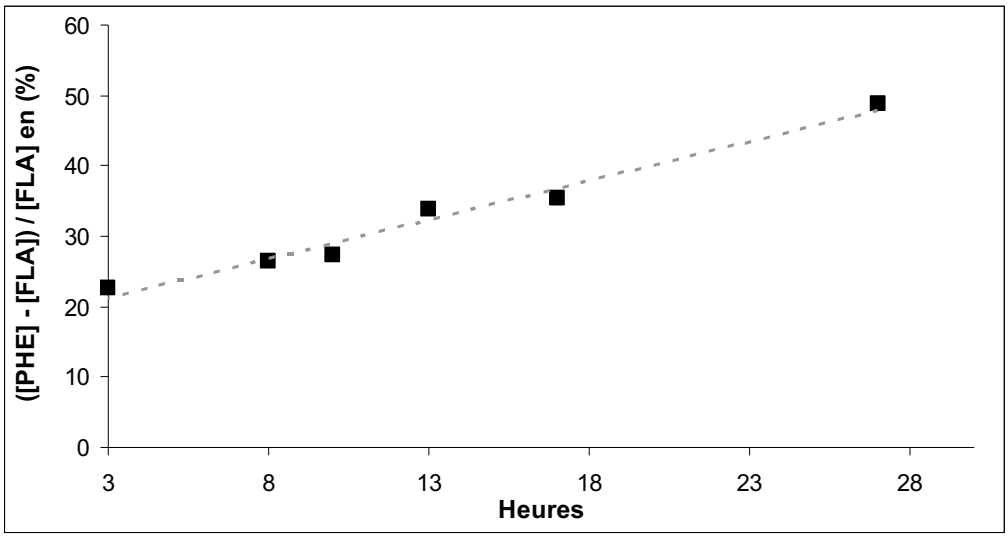


Figure 48 : Dégradation préférentielle du phénanthrène par rapport au fluoranthène au cours du temps.
Représentation en pourcentage par rapport au fluoranthène restant, de la différence entre la concentration restante de phénanthrène (PHE), et celle de fluoranthène (FLA) au cours du temps. Formule : $(Y = [([PHE](t) - [FLA](t))/[FLA](t)] \times 100)$.

Tableau 19 : Sondes dégénérées déterminées avec Metabolic Design pour chacun des 8 gènes considérés pour l'étude la souche *Sphingomonas paucimobilis* sp. EPA505.
Pour chaque gène, deux régions ont été identifiées et ciblées pour la détermination des sondes.
Nomenclature : **M** : A et C ; **K** : G et T ; **R** : A et G ; **W** : A et T ; **S** : G et C ; **Y** : C et T ; **V** : A, C et T ; **H** : A, C et T ; **D** : A, G et T ; **B** : G, T et C ; **I** : A, C, G et T.

Gène	Nom de la sonde	Séquence de la sonde	Nombre de sondes spécifiques	Positions sur le gène de référence
phnA1a	phnA1a_MD_A	GTITGYAAYTAYCAYGGITGGGT	256	294 – 316
	phnA1a_MD_B	CAYGARATHGARGTITGGACITA	384	957 – 979
phnA2a	phnA2a_MD_A	GARGAYATHCAYTAYTGGATGCC	48	123 – 145
	phnA2a_MD_B	GGICARGTITGGATGGARGAYCC	128	261 – 284
ahdA1c	ahdA1c_MD_A	GARTGYGTITAYCAYCARTGGGC	128	318 – 340
	ahdA1c_MD_B	GAYGCIGCIGAYAARCARGCITA	1024	771 – 793
ahdA2c	ahdA2c_MD_A	GAYGAYMGIYTIGARGARTGGCC	1024	081 – 103
	ahdA2c_MD_B	ATHGAYACIATGATGGTIMGICC	768	459 – 481
bphB	bphB_MD_A	AAYGTIGGIATHTGGGAYTWYAT	768	261 – 283
	bphB_MD_B	AAYBTIAARGGITAYTTYTTYGG	384	348 – 370
bphC	bphC_MD_A	CCITAYTTYATGCAYTGYAAYGA	128	558 – 580
	bphC_MD_B	TGGYTITGGGARTTYGGITGGGG	128	777 – 799
bphA3	bphA3_MD_A	ATHATHGARTGYCCITTYCAYGG	576	180 – 202
	bphA3_MD_B	ATHGAIGAYGGITGGGTITGYAT	768	279 – 302
ahdA4	ahdA4_MD_A	GCIAAYGTICCIGAYAAYTTYTT	1024	159 – 181
	ahdA4_MD_B	CARGARACITAYCARAAYGCIGC	512	867 – 889

Il semble, d'après ces résultats, que le phénanthrène soit préférentiellement dégradé, par rapport au fluoranthène (Figure 47). Le calcul de la différence de concentration de phénanthrène et de fluoranthène, au cours du temps montre une augmentation linéaire de cet écart. Cette relation linéaire (R^2 = 0,9778) permet donc d'émettre l'hypothèse d'une consommation préférentielle du phénanthrène lorsque les deux composés sont disponibles (Figure 48).

3. Evaluation de l'expression et caractérisation des gènes codant pour les enzymes de dégradation des HAP à l'aide de la biopuce ADN fonctionnelle

Nous nous sommes tout d'abord focalisés sur l'analyse de l'expression de huit gènes (*phnA1a, phnA2a, bphA3, ahdA4, bphB, bphC, ahdA1c* et *ahdA2c*) ce qui représente 8 048 sondes déduites de 16 sondes dégénérées (Tableau 19). Pour chaque gène, deux régions ont été identifiées pour la détermination des sondes. Les gènes *phnA1a, phnA2a, bphA3, ahdA4* codent pour les sous-unités constituant la dioxygénase responsable de l'attaque initiale des HAP. Le gène *bphB*, quant à lui, code pour une *cis*-dihydrodiol déshydrogénase impliquée dans la deuxième étape de dégradation de plusieurs HAP. Le gène *bphC* (isolée chez notre souche modèle) code pour une dihydroxynaphtalène dioxygénase impliquée dans le clivage en position *meta* de plusieurs dihydroxy-HAP. Nous nous sommes également focalisés sur deux sous-unités d'oxygénase (codées par *ahdA1c* et *ahdA2c*) caractérisées chez *Sphingomonas* sp. P2. Ces deux sous-unités forme une monooxygénase qui est impliquée dans la dégradation du phénanthrène, en transformant l'acide 1-hydroxy-2-naphthoïque en 1,2-dihydroxynaphtalène (Pinyakong *et al.*, 2003a; Cho *et al.*, 2005). Cette étape est spécifique de la voie de Evans, déjà décrite précédemment, et connue chez les espèces du genre *Sphingomonas*.

3.1. Caractérisation des niveaux d'expression génique

Une première biopuce a été hybridé avec des ARN extraits de la souche *Sphingomonas paucimobilis* sp. EPA 505 après 3 heures d'une culture avec le glucose, comme seule source de carbone et d'énergie. Les résultats obtenus montrent qu'une seule sonde ciblant le gène *ahdA2c* en région A présente un signal significatif (SNR'= 6,69 ± 2,01) (Tableau 20). Ce résultat provient probablement d'une hybridation croisée car les sondes devant cibler la région B de ce même gène n'ont pas répondu. De plus, les gènes codant pour les autres sous-unités de l'enzyme où intervient AhdA2c ne sont pas exprimés, comme la sous-unité de ferrédoxine, codée par *bphA3*. Nous ne pouvons cependant pas exclure que la

<u>Tableau 20 :</u> Evaluation de l'expression des gènes codant les enzymes dégradant les HAP chez la souche *Sphingomonas paucimobilis* sp. EPA505 par une approche de biopuces ADN fonctionnelle.

GLU : Hybridation de la biopuce ADN avec les ARN de la souche bactérienne cultivée sur glucose, PHE : de la biopuce ADN avec les ARN de la souche bactérienne cultivée sur phénanthrène, FLA : de la biopuce ADN avec les ARN de la souche bactérienne cultivée sur fluoranthène, PHE + FLA : de la biopuce ADN avec les ARN de la souche bactérienne cultivée sur un mélange de deux HAP (phénanthrène et fluoranthène).

	Nom du gène	***phnA1a***		***phnA2a***		***ahdA1c***		***ahdA2c***		***bphB***		***bphC***		***bphA3***		***ahdA4***	
	Région ciblée	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
	Nombre de sondes spécifiques total	256	384	48	128	128	1024	1024	768	768	384	128	128	576	768	1024	512
GLU	Nombre de sondes positives (SNR' > 3)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	SNR' médian le plus fort détecté dans les conditions définies	X	X	X	X	X	X	6,69 ± 2,01	X	X	X	X	X	X	X	X	X
PHE + FLA	Nombre de sondes positives (SNR' > 3)	1	2	0	1	3	1	2	1	4	1	1	1	3	1	0	0
	SNR' médian le plus fort détecté dans les conditions définies	18,32 ± 3,64	6,62 ± 0,31	X	22,64 ± 3,21	8,61 ± 1,59	9,93 ± 1,32	8,92 ± 1,52	16,26 ± 2,45	5,79 ± 1,73	4,09 ± 0,66	4,47 ± 0,30	4,54 ± 0,81	36,87 ± 7,83	9,79 ± 1,39	X	X
PHE	Nombre de sondes positives (SNR' > 3)	1	1	0	1	2	1	1	0	0	0	0	0	1	1	0	3
	SNR' médian le plus fort détecté dans les conditions définies	22,25 ± 13,53	10,02 ± 2,31	X	17,12 ± 3,00	4,11 ± 0,29	4,38 ± 2,75	21,75 ± 0,66	X	X	X	X	X	20,24 ± 5,06	20,00 ± 5,84	X	3,19 ± 0,40
FLA	Nombre de sondes positives (SNR' > 3)	1	3	0	1	1	0	2	2	2	0	0	1	1	1	0	0
	SNR' médian le plus fort détecté dans les conditions définies	14,50 ± 1,61	4,48 ± 1,60	X	8,82 ± 3,23	4,83 ± 1,39	X	16,37 ± 4,91	3,98 ± 0,03	3,37 ± 1,99	X	X	8,22 ± 1,16	15,61 ± 3,37	7,50 ± 2,03	X	X

région B de ce gène soit thermodynamiquement défavorable (présence de structures secondaires) pour l'hybridation avec les sondes.

Une seconde hybridation est réalisée avec les ARN extraits de la souche bactérienne, après 3 heures de culture en présence d'un mélange des deux HAP (phénanthrène et fluoranthène). Des signaux d'hybridation témoignant d'une expression génique ont pu être identifiés pour sept des huit gènes analysés (Tableau 20). En effet, il est important de noter qu'aucune sonde ciblant le gène *ahdA4* ne donne de signal supérieur au seuil de détection (SNR' > 3). Le gène *ahdA4*, bien que codant une enzyme potentiellement impliquée dans la dégradation des HAP, pourrait ne pas être exprimé, ou bien présenter un niveau d'expression trop faible, et donc en dessous du seuil de détection de la biopuce. Il faut également noter qu'aucune sonde ciblant la région A du gène *phnA2a* ne donne de signal positif, alors qu'une sonde de la région B donne un signal très important (SNR' = 22,64 ± 3,21). Des critères thermodynamiques défavorables pour la formation des duplex sondes/cibles pourraient expliquer ce résultat. De façon générale, les résultats d'hybridation montrent que peu de sondes ciblant une même région montrent un signal supérieur au seuil fixé, et ce pour chacun des gènes étudiés (Tableau 20). Certaines sondes donnent des signaux très élevés (*bphA3* SNR' = 36,87 ± 7,83, *phnA2a* SNR' = 22,64 ± 3,21, *phnA1a* SNR' = 18,32 ± 3,64), alors que les autres sondes ciblant la même région ne donnent généralement pas de signaux supérieurs au seuil défini. Ces résultats laissent donc supposer que les sondes présentant les SNR' les plus élevés s'apparient de façon parfaite à la séquence des gènes ciblés.

Les hybridations des ARN extraits après 3 heures de culture en présence de phénanthrène ou de fluoranthène comme seule source de carbone et d'énergie, donnent des résultats similaires (Tableau 20). Toutefois, l'expression du gène *ahdA4* est cette fois-ci détectée avec les ARN extraits en présence de phénanthrène. Etonnamment, aucune sonde ciblant les gènes *bphB* et *bphC* ne répond avec les ARN extraits en présence de phénanthrène seul. De plus, pour d'autres gènes, une des sondes ciblant l'une des deux régions étudiées ne répond plus avec les ARN extraits en présence d'un seul composé, alors qu'un signal positif était mesuré avec les ARN extraits en présence de deux composés. C'est le cas par exemple de la sonde ciblant la région B du gène *bphB*. Une analyse plus précise des images obtenues montre que ces sondes sont situées dans des zones présentant un fort bruit de fond local. Le SNR' de chacune des sondes étant calculé par rapport au bruit de fond local, on obtient alors des signaux dits « faux négatifs ».

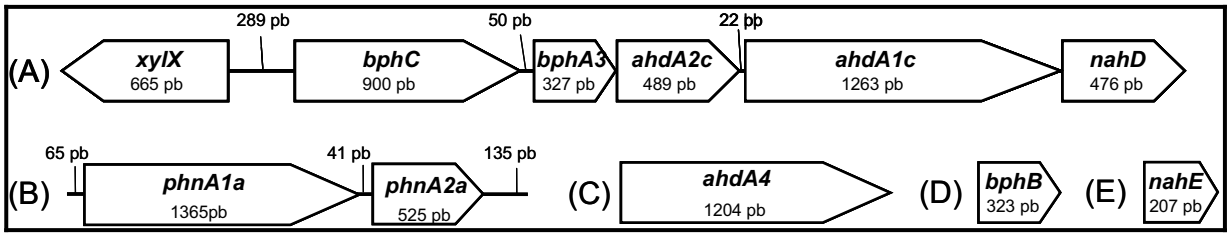


Figure 49 : Organisation génétique des cinq contigs (A, B, C, D et E) isolés pour la souche *Sphingomonas paucimobilis* sp. EPA505 des gènes codant les enzymes clés des voies de biodégradation des HAP.

La taille de chaque gène et les espaces intergéniques sont indiqués sur le schéma. (A) *xylX* : codant une sous-unité α de la toluène dioxygénase putative ; *bphC* : codant une dihydroxynaphtalène dioxygénase putative ; *bphA3* : codant une sous-unité de ferrédoxine putative ; *ahdA2c* : codant une petite sous-unité d'oxygénase putative ; *ahdA1c* : codant une grande sous-unité d'oxygénase putative ; *nahD* : codant une 2-hydroxychromène-2-carboxylate isomérase putative. (B) : *phnA1a* : codant une sous-unité α de la dioxygénase initiale putative ; *phnA2a* : codant une sous-unité β de la dioxygénase initiale putative. (C) : *ahdA4* : codant une ferrédoxine réductase putative. (D) : *bphB* : codant une *cis*-dihydrodiol déshydrogénase putative. (E) : *nahE* : codant une dihydroxybenzylpyruvate aldolase putative.

```
FM882255   1    ATGGCAGCAGTCACGGAACTCGGTTACCTTGGGTTGACCGTCACGAACCTCGATGCATGG  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   283  ATGGCAGCAGTCACGGAACTCGGTTACCTTGGGTTGACCGTCACGAACCTCGATGCATGG  342

FM882255   61   CGCAGTTATGCTGCCGAAGTGGCCGGCATGGAGGTTGTCGACGAGGGCGAAGGCGACCGC  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   343  CGCAGTTATGCTGCCGAAGTGGCCGGCATGGAGGTTGTCGACGAGGGCGAAGGCGACCGC  402

FM882255   121  CTCTACCTGCGCATGGACCAGTGGCATCATCGCATCGTGTTGCATGCCTCCGACTCCGAC  180
                ||||||||||||||||||||||||||||||||||||||||||||||| || ||||||||
AF259398   403  CTCTACCTGCGCATGGACCAGTGGCATCATCGCATCGTGTTGCATGCATCAGACTCCGAA  462

FM882255   181  GATCTTGCCTATCTGGGCTGGCGCGTTGCCGATCCGGTGGAATTCGACGCCATAGTGGCA  240
                ||| ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
AF259398   463  GATATTGCCTATCTGGGATGGCGCGTTGCCGATCCGGTGGAATTCGACGCCATAGTGGCA  522

FM882255   241  AAGCTGACCGCCGCCGGAATCTCCTTGACGGTGGCAAGCGAGGCCGAAGCTCGCGAGCGG  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   523  AAGCTGACCGCCGCCGGAATCTCCTTGACGGTGGCAAGCGAGGCCGAAGCTCGCGAGCGG  582

FM882255   301  CGCGTGCTCGGTCTTGCCAAGCTGGCTGATCCGGGTGGGAACCCCACCGAAATTTTTTAC  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   583  CGCGTGCTCGGTCTTGCCAAGCTGGCTGATCCGGGTGGGAACCCCACCGAAATTTTTTAC  642

FM882255   361  GGGCCGCAAGTCGACACCCACAAGCCTTTCCATCCTGGGCGCCCCATGTACGGAAAGTTC  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   643  GGGCCGCAAGTCGACACCCACAAGCCTTTCCATCCTGGGCGCCCCATGTACGGAAAGTTC  702

FM882255   421  GTGACCGGATCGGAAGGAATCGGGCATTGCATCCTGCGTCAGGACGATGTTCCCGCGGCG  480
                ||||||||||||||||||||||||||||||||||||||||||
AF259398   703  GTGACCGGATCGGAAGGAATCGGGCATTGCATCCTGCGTC--------------------  742

FM882255   481  GCGGCGTTCTACGGACTGCTGGGGCTGCGCGGGTCGGTCGAGTATCACTTGCAACTGCCC  540
                                              |||||    ||||||||||||||||||||
AF259398   743  ------------------------------GGTCG----AGTATCACTTGCAACTGCCC  767

FM882255   541  AACGGGATGGTGGCGCAGCCGTACTTCATGCACTGCAACGAGCGGCAGCATTCGGTCGCT  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   768  AACGGGATGGTGGCGCAGCCGTACTTCATGCACTGCAACGAGCGGCAGCATTCGGTCGCT  827

FM882255   601  TTCGGGCTTGGCCCGATGGAAAAGCGCATCAACCACCTGATGTTCGAATATACCGACCTC  660
                |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
AF259398   828  TTCGGGCTTGGCCCGATGCAAAAGCGCATCAACCACCTGATGTTCGAATATACCGACCTC  887

FM882255   661  GACGATCTCGGCCTCGCGCACGACATTGTGCGGGCGCGCAAGATCGACGTCGCGCTCCAG  720
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF259398   888  GACGATCTCGGCCTCGCGCACGACATTGTGCGGGCGCGCAAGATCGACGTCGCGCTCCAG  947

FM882255   721  CTCGGCAAGCACGCGAATGACCAGGCGCTGACCTTCTACTGCGCCAATCCGTCGGGCTGG  780
                |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
AF259398   948  CTCGGCAAGCACGCGAATGACCAGGCGCTGACCTTGTACTGCGCCAATCCGTCGGGCTGG  1007

FM882255   781  CTGTGGGAGTTCGGCTGGGGTGCCCGCAAGGCCCCGAGCCAGCAGGAATA  830
                ||||||||||||||| ||||||||||||||||||||||||||||||||||
AF259398   1008 CTGTGGGAGTTCGGGTGGGGTGCCCGCAAGGCCCCGAGCCAGCAGGAATA  1057
```

Figure 50 : Alignement partiel des deux séquences nucléiques du gène *bphC* de la souche *Sphingomonas paucimobilis* sp. EPA505.

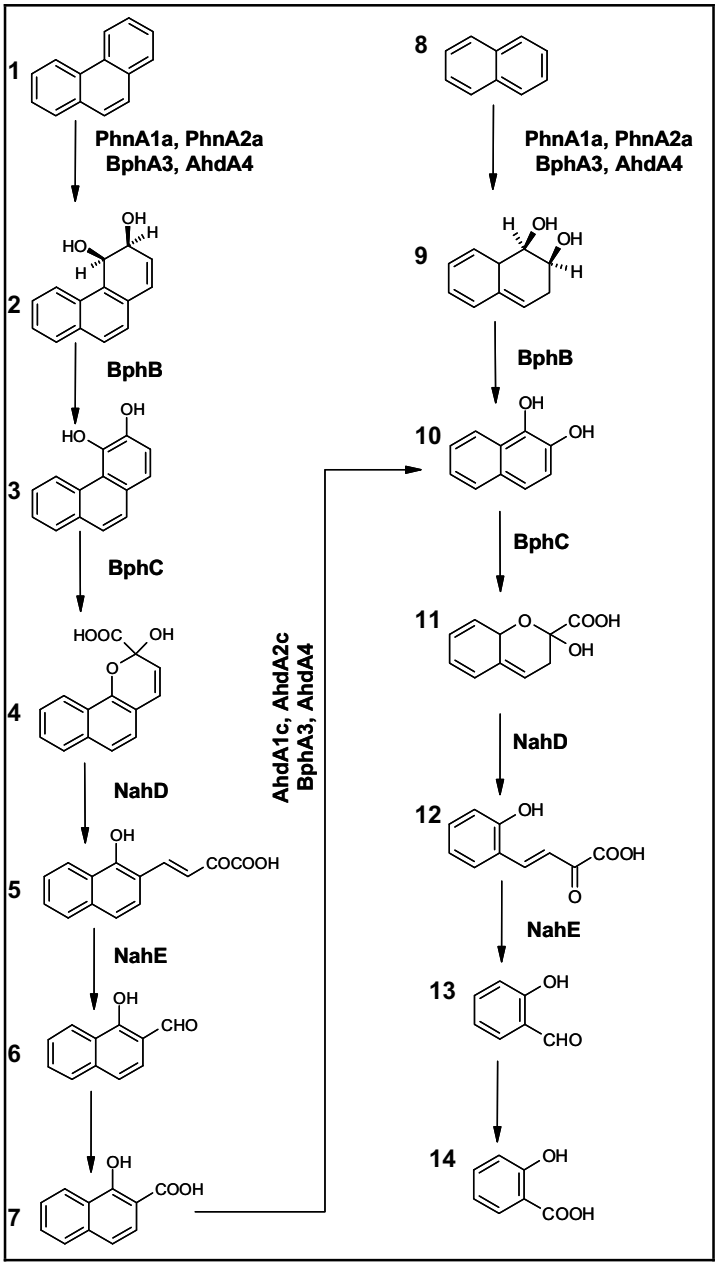
L'alignement est réalisé entre la séquence AF259398 (Story *et al.*, 2000), et celle isolée durant cette étude (FM882255).

En conclusion, les résultats d'hybridation montrent que l'expression des huit gènes est induite en présence de HAP, mais que les niveaux d'expression sont modulés en fonction du, ou des HAP présents comme source de carbone et d'énergie. Ainsi, d'après ces premiers résultats, le gène *ahdA4* semble plus spécifiquement exprimé en présence de phénanthrène seul. De plus, pour chaque région ciblée, et parmi la totalité des sondes spécifiques, il s'avère qu'une sonde spécifique particulière donne, dans la plupart des cas, un signal significativement plus élevé que les autres sondes ciblant la même région, et ce, quelque soit la source de HAP utilisée. Il est donc possible d'émettre l'hypothèse que ces sondes s'hybrident parfaitement avec les gènes ciblés de la souche EPA505. Afin de confirmer cette hypothèse, une stratégie d'amplification et de clonage, séquençage a été mise en place pour isoler les gènes étudiés chez *Sphingomonas paucimobilis* sp. EPA505.

3.2. Isolement et caractérisation des gènes codant les enzymes de dégradation des HAP chez la souche Sphingomonas paucimobilis EPA 505.

La stratégie d'amplification par PCR, de clonage et de séquençage (voir Matériel et méthodes) des gènes étudiés a permis la caractérisation de cinq contigs de 4,47 kpb, 2,13 kpb, 1,20 kpb, 0,32 kpb et 0,20 kpb (respectivement FM882255, FM882254, FM882253, FN552592 et FN552593) (Figure 49).

Le premier contig de 4,47 kbp présente 6 cadres de lecture ouverts (ORF) putatifs (Figure 49, A). Le premier ORF, dont la séquence est partielle, et présentant 90 % d'identité nucléique avec EF151283.1 de *Sphingomonas yanoikuyae* B1, code pour un polypeptide 94 % identique à l'enzyme XylX (ABM79785). Cette enzyme, correspondant à une sous-unité d'oxygénase, est impliquée dans la dégradation du toluène et du xylène chez la souche B1 (Pinyakong *et al.*, 2003a; Desai *et al.*, 2008). Les cinq autres ORF sont localisés sur l'autre brin et sont séparés par de courtes distances intergéniques (0 à 50 pb), laissant supposer une organisation en opéron. Ces cinq ORF présentent tous une identité élevée avec des gènes déjà isolés de la souche *Sphingomonas* sp. P2. Le premier ORF (montrant 97 % d'identité nucléique avec AB091692.1 isolée chez *Sphingomonas* sp. P2) code pour un polypeptide 98 % identique à une enzyme nommée BphC (BAC65429), qui est une dihydroxynaphtalène dioxygénase connue pour dégrader divers dihydroxy-HAP. Cependant, la séquence nucléique de cet ORF présente seulement une identité de 93 % avec le gène *pbhB* déjà isolé chez la souche EPA505 (AF259397 et AF259398) (Story *et al.*, 2000). Néanmoins, ce pourcentage d'identité assez faible entre les deux séquences nucléiques peut s'expliquer par la présence d'une importante délétion (de 60 pb) au sein de la séquence de *pbhB* (Figure 50). Le second



<u>Figure 51 :</u> Etapes enzymatiques où sont impliquées les protéines putatives codées par les gènes identifiés chez *Sphingomonas paucimobilis* sp. EPA505.

Les flèches sans lettres peuvent représenter une ou plusieurs étapes enzymatiques.

Composés : (**1**) Phénanthrène ; (**2**) *cis*-3,4-phénanthrène dihydrodiol ; (**3**) 3,4-dihydroxyphénanthrène ; (**4**) 2-hydroxybenzo(*h*)chromène-2-carboxylate ; (**5**) 4-(1-hydroxynapht-2-yl)-2-oxobut-3-énoate ; (**6**) 1-hydroxy-2-naphtaldéhyde ; (**7**) 1-hydroxy-2-napthoate ; (**8**) naphtalène ; (**9**) *cis*-1,2-naphtalène dihydrodiol ; (**10**) 1,2-dihydroxynaphtalène ; (**11**) 2-hydroxychromène-2-carboxylate; (**12**) *trans-o*-hydroxybenzylidène pyruvate ; (**13**) salicylaldéhyde ; (**14**) salicylate.

Enzymes putatives : Dioxygénase initiale (PhnA1a et PhnA2a correspondant respectivement aux sous-unités d'oxygénase α et β, BphA3 à la sous-unité de ferrédoxine et AhdA4 à la ferrédoxine réductase) ; (b) dihydrodiol déshydrogénase (BphB) ; (c) dihydroxynaphtalène dioxygénase (BphC) ; (d) 2-hydroxychromène-2-carboxylate isomérase (NahD) ; (e) dihydroxybenzylpyruvate aldolase (NahE) ; (f) salicylate oxygénase (AhdA1c et AhdA2c correspondant respectivement aux sous-unités α et β d'oxygénase, BphA3 à la ferrédoxine et AhdA4 à la ferrédoxine réductase).

des cinq ORF (présentant 89 % d'identité nucléique avec AB091692.1 isolée chez *Sphingomonas* sp. P2) code pour un polypeptide 90 % identique à une sous-unité de ferrédoxine, nommée BphA3 (BAC65428). Cette sous-unité est connue pour être un intermédiaire du transfert des électrons, de la ferrédoxine réductase, au complexe de la dioxygénase et impliquée dans diverses étapes de dégradation des HAP (Story *et al.*, 2000; Pinyakong *et al.*, 2003b). Les deux gènes situés après *bphA3* codent pour deux polypeptides respectivement identiques à 88 % et 95 % à AhdA2c (BAC65427) et AhdA1c (BAC65426) (et présentant respectivement 87 % et 93 % d'identité nucléique avec AB091692.1 de *Sphingomonas* sp. P2). Ces deux enzymes sont deux sous-unités d'oxygénase impliquées dans la conversion du salicylate en catéchol, mais aussi dans la formation du 1,2-dihydroxynaphtalène à partir de l'acide 1-hydroxy-2-naphtoïque (Pinyakong *et al.*, 2003b; Cho *et al.*, 2005). Le dernier ORF (présentant 92 % d'identité nucléique avec AB091692.1 isolée chez *Sphingomonas* sp. P2), partiellement séquencé, code pour un polypeptide de 146 acides aminés dont la séquence est identique à 91 % à une 2-hydroxychromène-2-carboxylate isomérase, nommée NahD (BAC65425). Cette enzyme est connue pour être impliquée dans la dégradation du naphtalène et du phénanthrène (Kim *et al.*, 1997). Tous ces polypeptides interviennent à différentes étapes de dégradation des HAP (Figure 51).

Les deux ORF identifiés au sein du second contig de 2,13 kpb (FM882254), présentent respectivement 99 % et 100 % d'identité nucléique avec AJ633551 de *Sphingomonas* sp. CHY-1, et codent pour deux polypeptides de 455 et 175 résidus. Ces deux polypeptides sont respectivement identiques à 99 % et à 100 % aux sous-unités α et β de la dioxygénase initiale isolées chez *Sphingomonas* sp. CHY-1, nommées PhnA1a et PhnA2a (CAG17576 et CAG17577) (Figure 49, B). Ces deux sous-unités sont connues pour être impliquées dans l'attaque initiale de plusieurs HAP chez plusieurs souches de *Sphingomonas* (Figure 51) (Demaneche *et al.*, 2004; Jakoncic *et al.*, 2007a, b).

Le troisième contig de 1,20 kpb (FM882253) ne comporte qu'un seul ORF partiel (présentant 92 % d'identité nucléique avec AB091693.1 de *Sphingomonas* sp. P2), et montrant une identité de 95 % avec la séquence protéique d'une ferrédoxine réductase, nommée AhdA4, isolée chez *Sphingomonas* sp. P2 (BAC65450). Cette réductase est impliquée dans les mêmes étapes de transfert d'élection que BphA3 (Figures 49, C et 51) (Pinyakong *et al.*, 2003a).

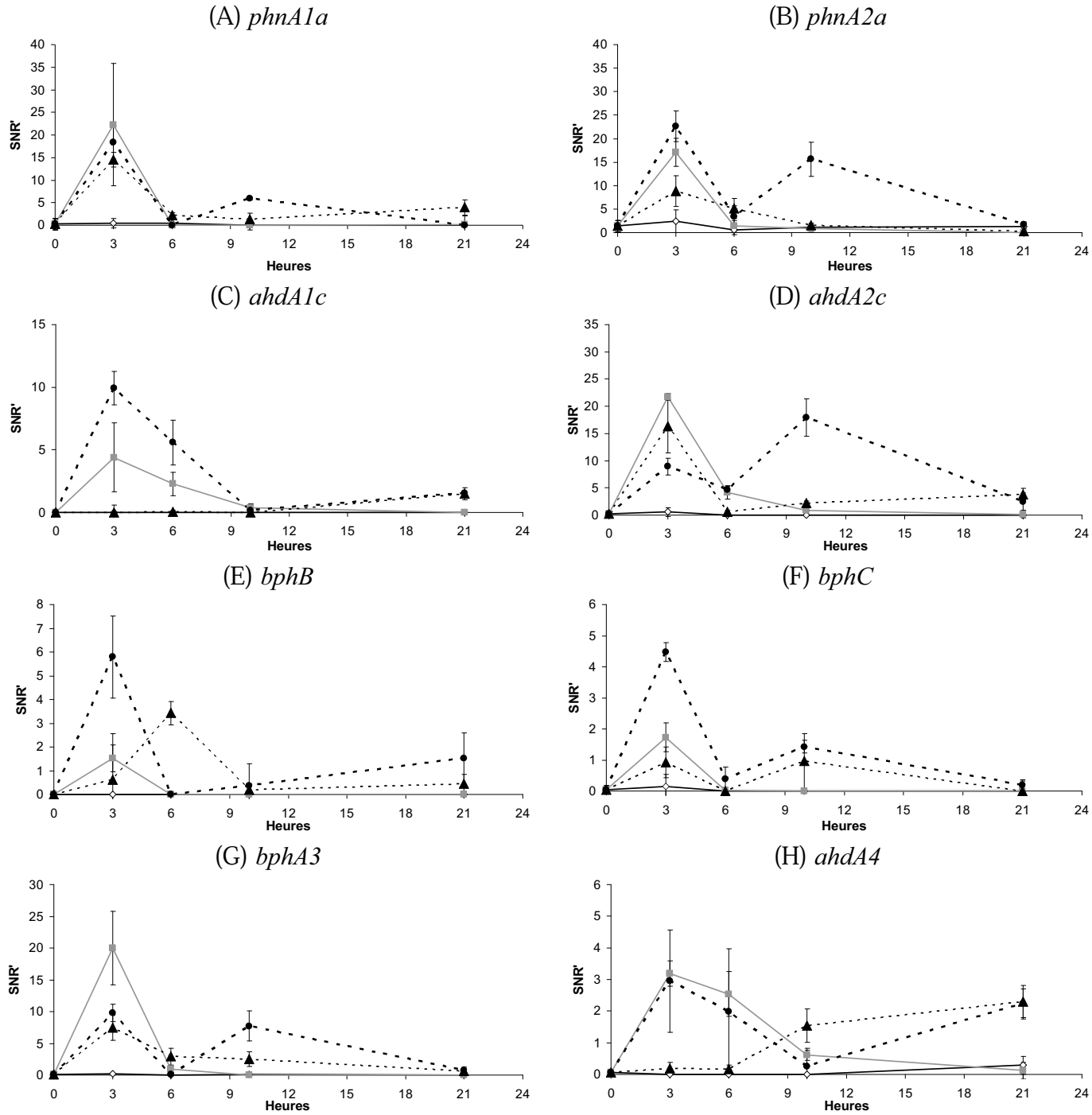
Le quatrième contig de 0,32 kpb (FN552592) (présentant 96 % d'identité nucléique avec EF151283.1 de *Sphingomonas yanoikuyae* B1), code pour un polypeptide partiel de 107 acides aminés, 97 % identique à une *cis*-dihydrodiol déshydrogénase nommée BphB

(ABM79802), isolée chez *Sphingomonas yanoikuyae* B1 (Figure 49, D). Cette enzyme est connue pour être impliquée dans la conversion de plusieurs dihydrodiols en leurs composés dihydroxylés correspondants (Figure 51) (Pinyakong *et al.*, 2003a). Enfin, le dernier contig de 0,20 kpb (FN552592) (montrant 94 % d'identité nucléique avec AB091693.1 de *Sphingomonas* sp. P2), encode un polypeptide partiel de 68 acides aminés, 92 % identique à une *trans*-*o*-hydroxybenzylidène-pyruvate hydratase-aldolase appelée NahE (BAC65452) isolée chez *Sphingomonas* sp. P2 (Figure 49, E). Cette enzyme est connue pour être impliquée dans la conversion du 4-(1-hydroxynapht-2-yl)-2-oxobut-3-énoate en 1-hydroxy-2-naphtaldéhyde (Habe et Omori, 2003) et spécifique des voies de dégradation du phénanthrène et du naphtalène (Figure 51) (Story *et al.*, 2000). La séquence partielle du gène *nahE* obtenue ne nous permet de vérifier notre hypothèse, car trop courte, les régions ciblées par les sondes déterminées n'ont donc pas été isolées.

3.3. Comparaison des résultats de biopuces ADN et de séquençage

Les sondes exploratoires développées sans *a priori* sur les séquences des gènes codant les enzymes de dégradation des HAP montrent qu'elles sont capables de détecter leur expression lorsque la souche EPA505 est mise en culture avec le phénanthrène et le fluoranthène. L'analyse des résultats laisse également supposer que les sondes montrant les SNR' les plus élevés sont celles possédant les séquences complémentaires aux séquences des gènes de la souche étudiée. Cette hypothèse a pu être validée dans la majorité des cas par la comparaison entre les séquences des sondes déterminées et celles des gènes isolées de la souche EPA505. En effet, les résultats obtenus avec les ARN extraits en présence d'un mélange de phénanthrène et de fluoranthène montrent que parmi les treize sondes donnant les plus hauts SNR', neuf ont une séquence 100 % identique aux gènes de la souche EPA505. C'est ainsi le cas pour les deux sondes (régions A et B) ciblant les gènes *bphA3* et *ahdA2c*, et pour une des deux sondes ciblant les gènes *phnA1a*, *phnA2a*, *ahdA1c*, *bphB* et *bphC*. Ces résultats sont de plus confirmés avec les ARN extraits en présence d'un seul HAP. En effet, avec les ARN extraits en présence uniquement de fluoranthène pour la croissance de la souche bactérienne, sept sondes sur les dix donnant les plus hauts SNR' présentent une séquence 100 % identique aux gènes de la souche EPA505, et sept sondes sur neuf, dans le cas où le phénanthrène est utilisé comme seule source de carbone et d'énergie.

Ces résultats valident la stratégie exploratoire de la biopuce ADN fonctionnelle, assurant ainsi l'identification de gènes sans *a priori* sur leurs séquences. Par cette approche, il est donc possible de déterminer la séquence génique ciblée par la sonde. La très bonne



<u>Figure 52 :</u> Cinétiques d'expression des gènes codant pour des enzymes de dégradation des HAP chez la souche *S. paucimobilis* sp. EPA505 par l'approche biopuce fonctionnelle.

Sources carbonées utilisées durant la croissance bactérienne : Losange blanc : glucose, carré gris : phénanthrène, triangle noir : fluoranthène, rond noir : phénanthrène et fluoranthène. Gènes étudiés : *phnA1a* (codant une sous-unité α de la dioxygénase initiale), *phnA2a* (sous-unité β de la dioxygénase initiale), *ahdA1c* (grande sous-unité d'oxygénase), *ahdA2c* (petite sous-unité d'oxygénase), *bphB* (*cis*-dihydrodiol déshydrogénase), *bphC* (dihydroxynaphtalène dioxygénase), *bphA3* (sous-unité de ferrédoxine) et *ahdA4* (ferrédoxine réductase). Les barres d'erreur représentent l'écart-type des SNR' mesurés pour chacun des trois réplicats des sondes étudiées.

spécificité de ces sondes est également confirmée par les résultats obtenus avec les ARN extraits suite à la culture de la souche avec du glucose, où une seule sonde sur 8 048 montre un signal positif.

Les valeurs de SNR' obtenues montrent cependant une grande variabilité en fonction des sondes, et ce pour chaque gène. C'est le cas, par exemple, du gène *phnA2a*, où aucune des sondes ciblant la région A ne répond, alors que la région B donne un signal important (Tableau 20). Une des hypothèses émises est que certaines régions des cibles peuvent limiter l'accessibilité d'appariement des cibles du fait de la présence de structures secondaires. Une autre hypothèse serait la faible sensibilité de certaines sondes. Malgré ces variations de sensibilité, il est cependant important de s'assurer que l'approche biopuce ADN permet de réaliser des suivis semi quantitatif de l'expression des gènes ciblés. Pour cela, les biopuces ADN ont été hybridées avec des ARN extraits au cours d'une cinétique de dégradation du phénanthrène, du fluoranthène et du mélange de ces deux HAP, par la souche EPA505. La confirmation des résultats sera réalisée par une approche de PCR quantitative sur les mêmes cinétiques.

3.4. Cinétique d'expression des gènes codant les enzymes de dégradation des HAP

Le suivi de l'expression des gènes a été réalisé avec les ARN extraits de la souche bactérienne aux temps 0, 3, 6, 10 et 21 heures des différentes cultures (glucose, phénanthrène, fluoranthène, phénanthrène + fluoranthène). Il est important de noter que, par rapport aux résultats décrits précédemment, nous ne nous sommes intéressés qu'aux sondes les plus efficaces et présentant une séquence identique à celles des gènes de la souche bactérienne étudiée. Enfin, l'expression de chacun des gènes étudiés a été quantifiée par une approche de PCR quantitative sur un plus grand nombre de points (aux temps 0, 3, 6, 8, 10, 13, 17, 21, 24 et 27 heures) afin de mieux appréhender les cinétiques d'expression.

3.4.1. Suivis d'expression génique par une approche de biopuce ADN

Les hybridations sont tout d'abord réalisées en utilisant les ARN extraits de la souche modèle, en culture avec du glucose comme seule source carbonée. Les résultats obtenus pour les sondes considérées ne montrent aucun signal positif tout au long de la cinétique de croissance (après 0, 3, 6, 10 et 21 heures de culture) (Figure 52). Cela démontre que les gènes étudiés ne sont pas induits au cours de la croissance sur glucose.

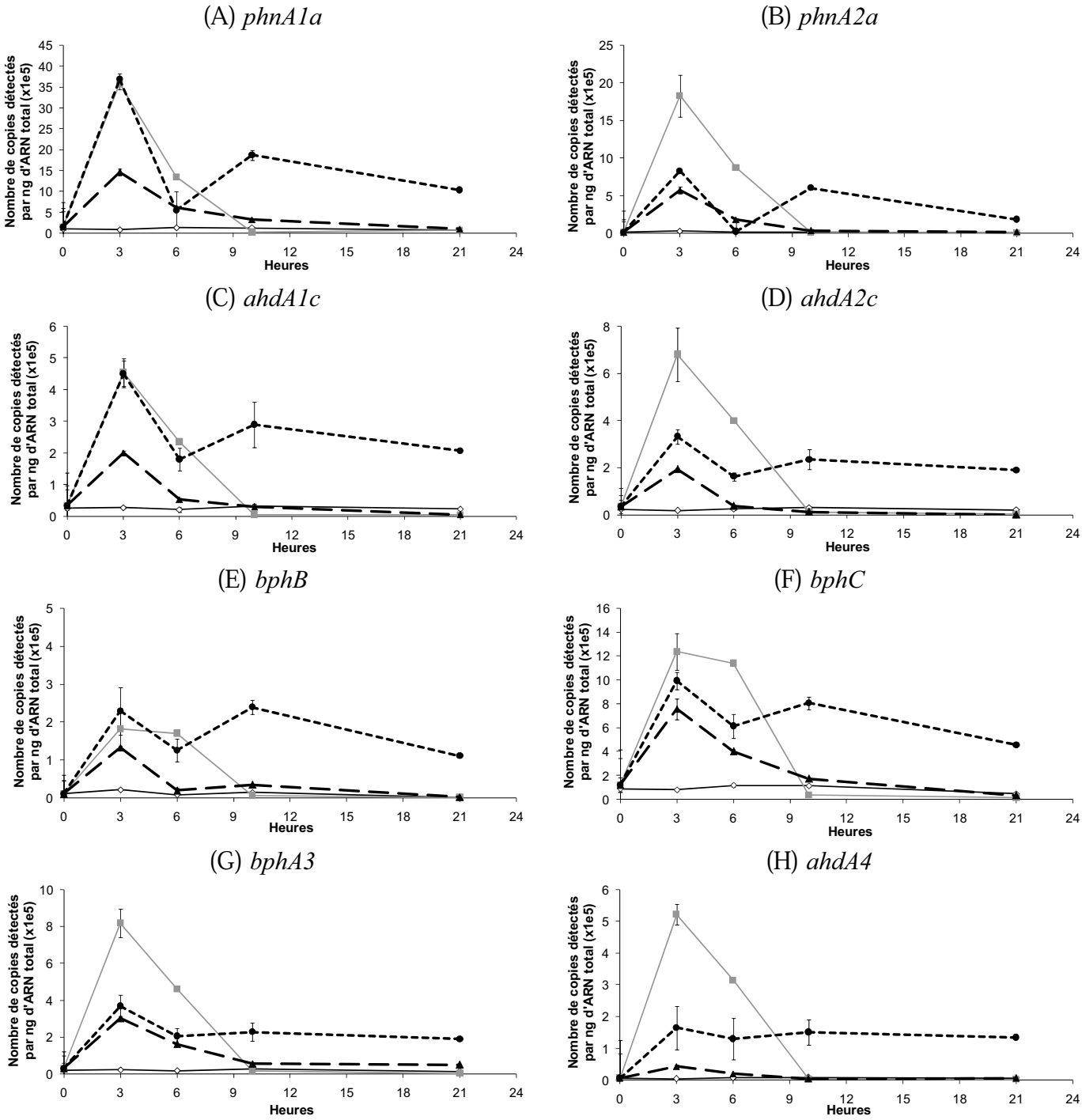
Dans le cas de la cinétique réalisée en présence de phénanthrène, seules les hybridations réalisées avec les ARN extraits après 3 heures de culture permettent d'obtenir de

forts signaux (Figure 52). Par exemple, un SNR' = 22,25 ± 13,53 est défini pour la sonde ciblant la région A du gène *phnA1a*. De même, d'autres sondes ciblant les gènes *phnA2a, ahdA1c, ahdA2c* ou *bphA3* répondent de manière importante (Figures 52 B, C, D et G). Avec ce substrat, l'expression du gène *ahdA4* est également détectée (SNR' = 3,19 ± 0,40). Toutefois, le signal obtenu est à la limite du seuil significatif (SNR' > 3). Les hybridations avec les ARN extraits après 6, 10 et 21 heures de culture montrent des SNR' pour les sondes considérées plus faibles que ceux obtenus après 3 heures (Figure 52).

Pour les ARN extraits de la culture en présence de fluoranthène seul, les sondes spécifiques aux gènes *phnA1a, phnA2a, ahdA2c, bphC* et *bphA3* de la souche EPA505 donnent des signaux supérieurs au seuil fixé avec les ARN extraits après 3 heures de culture (Figure 52). Néanmoins, aucun signal positif n'est détecté pour les sondes 100 % identiques aux séquences des gènes *bphB* et *ahdA1c*. Les valeurs des SNR' obtenues pour les ARN extraits après 6 heures de culture diminuent par rapport à celles obtenues après 3 heures, excepté pour *bphB*. En effet, un signal positif (SNR' = 3,43 ± 0,70) est détecté pour la sonde spécifique de la région A du gène *bphB* après 6 heures de culture (Figure 52, E). Enfin, les hybridations avec les ARN extraits après 10 et 21 heures de culture en présence de fluoranthène seul présentent des SNR' comparables à ceux détectés avec les ARN extraits avec du glucose comme seule source carbonée.

Enfin, les hybridations réalisées avec les ARN extraits des cultures réalisées en présence d'un mélange des deux composés permettent de mettre en évidence un profil d'expression au cours du temps différent de ceux précédemment décrits (Figure 52). Ainsi, après 3 heures de culture, des signaux positifs importants (approximativement de la même intensité que pour le phénanthrène seul) sont détectés pour les gènes *phnA1a, phnA2a, ahdA1c, ahdA2c, bphB, bphC* et *bphA3*. Par exemple, un SNR' = 18,32 ± 3,64 est mesuré pour la sonde ciblant la région A du gène *phnA1a* (Figure 52, A). Seul le gène *ahdA4* n'est pas détecté de façon significative (Figure 52, H) avec les sondes considérées. Ensuite, après être passés en dessous du seuil de détection pour le temps 6 heures de la cinétique, les signaux témoignent de nouveau d'une expression significative des gènes au temps 10 heures, mais avec cependant des valeurs de SNR' plus faibles que celles obtenues avec les ARN extraits après 3 heures de culture. Enfin, avec les ARN extraits après 21 heures de culture, les signaux sont situés à des valeurs inférieures au seuil fixé (SNR' > 3).

En résumé, ces résultats montrent une forte expression de la plupart des gènes étudiés en présence des différents HAP. En effet, seul le gène *ahdA4* ne semble exprimé qu'en présence de phénanthrène. Ces résultats peuvent laisser supposer que les sondes développées



<u>Figure 53 :</u> Cinétiques d'expression des gènes codant pour des enzymes de dégradation des HAP chez la souche *S. paucimobilis* sp. EPA505 par l'approche de PCR quantitative. Sources carbonées utilisées durant la croissance bactérienne : Losange blanc : glucose, carré gris : phénanthrène, triangle noir : fluoranthène, rond noir : phénanthrène et fluoranthène. Gènes étudiés : *phnA1a* (codant une sous-unité α de la dioxygénase initiale), *phnA2a* (sous-unité β de la dioxygénase initiale), *ahdA1c* (grande sous-unité d'oxygénase), *ahdA2c* (petite sous-unité d'oxygénase), *bphB* (*cis*-dihydrodiol déshydrogénase), *bphC* (dihydroxynaphtalène dioxygénase), *bphA3* (sous-unité de ferrédoxine) et *ahdA4* (ferrédoxine réductase). Les barres d'erreur représentent l'écart-type des SNR' mesurés pour chacun des trois réplicats des sondes étudiées.

pour ce gène ne sont pas assez sensibles pour assurer la détection des ARN messagers. Enfin, le niveau d'expression des autres gènes semble lié à la nature du ou des composé présent(s).

3.4.2. Suivi d'expression génique par une approche de PCR quantitative

Pour valider les résultats obtenus par la biopuce ADN sur la modulation de l'expression des gènes au cours des cinétiques de dégradation des HAP, une approche de PCR quantitative a été mise en place pour quantifier le nombre de transcrits spécifiques de chacun des huit gènes étudiés, et ce sur les mêmes cinétiques de biodégradation. Les niveaux d'expression détectés sont très faibles pour la culture témoin (en présence de glucose) et restent stables tout au long de la cinétique (Figure 53 et Annexe 6).

En présence de phénanthrène comme seule source de carbone et d'énergie, le suivi du nombre de transcrits révèle que tous les gènes sont surexprimés, d'un facteur 10 pour *bphC* à un facteur 112 pour *phnA2a*, par rapport aux niveaux mesurés en présence de glucose (Figure 53 et Annexe 6). Les profils obtenus montrent que le nombre de transcrits est le plus important après 3 heures de culture ($1{,}24 \times 10^6$ - $8{,}18 \times 10^5$ - $6{,}81 \times 10^5$ - $4{,}55 \times 10^5$ - $3{,}58 \times 10^6$ - $1{,}83 \times 10^6$ - $5{,}22 \times 10^5$ et $1{,}82 \times 10^5$ molécules par ng d'ARN total respectivement pour *bphC*, *bphA3*, *ahdA2c*, *ahdA1c*, *phnA1a*, *phnA2a, ahdA4* et *bphB*), puis ce nombre baisse rapidement pour rejoindre un niveau basal après 8 heures de culture.

Le même profil est visible pour les cinétiques en présence de fluoranthène, mais avec des niveaux d'expression plus faibles que ceux mesurés en présence de phénanthrène pour la plupart des gènes (Figure 53 et Annexe 6). Le nombre de transcrits le plus élevé est également détecté après 3 heures de culture ($7{,}57 \times 10^5$ - $3{,}02 \times 10^5$ - $1{,}93 \times 10^5$ - $2{,}00 \times 10^5$ - $1{,}47$ x 10^6 - $5{,}47 \times 10^5$ - $4{,}28 \times 10^4$ et $1{,}32 \times 10^5$ molécules par ng d'ARN total respectivement pour *bphC*, *bphA3*, *ahdA2c*, *ahdA1c*, *phnA1a*, *phnA2a, ahdA4* et *bphB*), puis, après 10 heures de culture, ce nombre diminue pour rejoindre des valeurs basales proches de celles obtenues en présence de glucose.

Enfin, pour la culture réalisée avec un mélange des deux HAP, les transcrits sont également fortement représentés après 3 heures de culture, avec des concentrations situées entre celles mesurées pour le phénanthrène, et celles pour le fluoranthène (Figure 53 et Annexe 6). Puis, malgré une baisse importante enregistrée après 6 heures pour tous les gènes étudiés, une nouvelle augmentation du nombre de transcrits se produit après 8 heures (pour *bphC, bphA3, phnA1a, phnA2a* et *bphB*) ou 10 heures (pour *ahdA1c, ahdA2c* et *ahdA4*). Enfin, les profils montrent de nouveau une baisse de l'expression des gènes, rejoignant le niveau détecté en présence de glucose après 17 ou 21 heures de culture.

3.4.3. Comparaison des résultats obtenus avec les approches de biopuce ADN et de PCR quantitative

Les résultats obtenus par PCR quantitative confirment les profils d'expression obtenus par la biopuce pour chacune des cinétiques de dégradation étudiées. En effet, les deux approches donnent des profils d'expression similaires (Figures 52 et 53). Ces résultats démontrent donc l'efficacité des sondes développées avec Metabolic Design. En effet, les sondes développées, dans la majorité des cas permettent de réaliser un suivi semi quantitatif de l'expression des gènes étudiés.

Toutefois, pour le gène *ahdA4*, l'approche de PCR quantitative montre que ce dernier est faiblement exprimé en présence de HAP. Les sondes ciblant *ahdA4* ne sont donc pas aussi sensibles, même si nous observons le même profil proche du seuil de détection significatif. En effet, le gène *ahdA4* est uniquement détecté à la limite du seuil défini (SNR' = 3,19 ± 0,40) avec les ARN extraits après trois heures de culture, en présence de phénanthrène.

4. Analyse de l'expression des autres gènes ciblés par la biopuce ADN

La biopuce développée permet d'étudier l'expression d'un total de 40 gènes. Suite à la validation de l'approche portant sur les 8 gènes décrits précédemment, les résultats des 32 autres gènes ont été analysés avec les ARN extraits après trois heures de culture. Seuls les résultats les plus significatifs seront détaillés ci-après.

Les gènes *ahdA1d* et *ahdA2d* codant pour deux sous-unités d'oxygénase, potentiellement impliquées dans la dégradation des HAP (Romine *et al.*, 1999b). Les résultats pour *ahdA1d* montrent des niveaux d'expression importants en présence de phénanthrène ou de fluoranthène avec des SNR' respectivement de 13,11 ± 0,19 et 15,85 ± 1,74. Une importante expression de ce gène, en présence de ces mêmes composés, avait été obtenue chez une souche proche nommée *Sphingomonas* sp. P2 (Pinyakong *et al.*, 2003b; Cho *et al.*, 2005). L'hybridation réalisée avec les ARN extraits de la souche bactérienne EPA505 en présence des deux composés, permet d'obtenir un signal encore plus important pour *ahdA1d* (SNR'= 35,80 ± 5,87). De façon surprenante, les signaux obtenus pour le gène *ahdA2d* ne montre qu'une expression en présence de phénanthrène seul (SNR'= 11,29 ± 1,51), alors qu'il est généralement admis que ces sous-unités sont co-exprimées. Les signaux obtenus sont en effet plus faibles avec les ARN extraits à partir de la culture en présence des deux composés (SNR' = 5,81 ± 1,04), ou en présence de fluoranthène (SNR' = 1,05 ± 0,45).

Pour le gène *xylE*, (codant une catéchol-2,3-dioxygénase impliquée dans la voie de clivage *meta* du catéchol), la sonde xylE_MD_B_0331 permet de mettre en évidence une surexpression de ce gène au cours des 3 cinétiques (SNR' = 6,77 ± 1,73; SNR' = 5,62 ± 0,44 et SNR' = 39,96 ± 4,95 respectivement obtenus avec les ARN extraits en présence de phénanthrène, de fluoranthène et d'un mélange des deux composés). D'après les résultats de la sonde bphK_MD_B_0035 ciblant le gène *bphK* (codant une glutathione-S-transférase), ce dernier est exprimé uniquement en présence du mélange des deux composés (SNR' = 13,78 ± 0,86).

Le gène *xylC* (codant une benzaldéhyde déshydrogénase, connue pour être impliquée dans la dégradation de molécules monoaromatiques comme le toluène (Romine *et al.*, 1999b) semble également être exprimé avec les ARN extraits en présence d'un mélange des deux composés (SNR' = 11,98 ± 1,28 pour la sonde xylC_MD_A_0468). Les SNR' mesurés dans les autres conditions sont proches ou en dessous du seuil fixé : entre 1,73 ± 1,90 et 3,26 ± 0,98 avec les ARN extraits en présence de phénanthrène ou de fluoranthène seul).

Dans le cas du gène *nahD* (codant la 2-hydroxychromène-2-carboxylate isomérase spécifiquement impliquée dans la voie de dégradation du naphtalène (Kim *et al.*, 1997)), bien qu'aucun signal positif ne soit détecté en présence de fluoranthène, un SNR' supérieur à la valeur seuil est déterminée en présence de phénanthrène (SNR'= 7,61 ± 1,13), ou en présence du mélange des deux HAP (SNR'= 10,47 ± 2,41). L'expression de ce gène serait donc spécifiquement induite en réponse à la présence de phénanthrène.

L'expression du gène *bphR*, codant un régulateur intervenant potentiellement dans les voies de régulation des gènes impliquées dans la dégradation des HAP (Romine *et al.*, 1999a) n'est visible qu'avec les ARN extraits en présence de glucose (SNR' = 13,98 ± 0,90 avec la sonde bphR_MD_A_0594). De plus, une expression proche du seuil défini (SNR' = 3,20 ± 1,37) n'est visible qu'avec les ARN extraits en présence d'un mélange de HAP. Au vu de ces résultats, cette protéine pourrait donc réguler négativement l'expression des gènes codant des protéines impliquées dans la biodégradation des HAP.

Enfin, l'expression du gène nommé Orf597, codant un régulateur putatif de la famille MucR, pouvant intervenir dans les voies de régulation des gènes impliquées dans la dégradation des HAP est induite. En effet, la sonde orf597_MD_A_0154 montre une surexpression de ce gène en réponse à la présence de phénanthrène (SNR' = 7,64 ± 4,75) ou des deux HAP (SNR' = 4,39 ± 0,14). Ces résultats nous permettent d'émettre l'hypothèse que le produit de ce gène joue un rôle potentiel dans la régulation des gènes impliqués dans la dégradation du phénanthrène.

5. Conclusion

Ce travail a permis d'évaluer l'efficacité (spécificité et sensibilité) des sondes déterminées avec Metabolic Design et de démontrer leur aspect exploratoire. Ainsi, la biopuce ADN développée a d'abord été utilisée pour évaluer l'expression de huit gènes codant des enzymes impliquées dans la dégradation des HAP étudiés.

Les résultats obtenus montrent les excellentes sensibilités et spécificités des sondes développées afin d'appréhender toute la diversité génique. En effet, les résultats soulignent le fait que les séquences des sondes donnant les SNR' les plus élevés sont généralement parfaitement complémentaires aux séquences des gènes ciblés. De même, ces résultats montrent que les sondes sont suffisamment sensibles pour réaliser un suivi semi quantitatif des gènes étudiés, même pour certains gènes faiblement exprimés (comme *bphB*). En effet, les résultats obtenus par PCR quantitative confirment les profils cinétiques mesurés avec les sondes déterminées pour la majorité des gènes étudiés.

Néanmoins, certaines sondes ne semblent pas assurer la détection des transcrits présents au sein de l'échantillon hybridé. Les contraintes thermodynamiques (Pozhitkov *et al.*, 2007; Mueckstein *et al.*, 2010) non estimées et vérifiées par Metabolic Design peuvent potentiellement expliquer ces comportements, tout comme les variations d'accessibilité des différentes régions ciblées au sein d'un même transcrit (Royce *et al.*, 2007). C'est pourquoi, il est préférable de réaliser tout d'abord une validation expérimentale de la spécificité et de la sensibilité des sondes développées, à l'aide d'un modèle simple.

Nos résultats montrent également une expression différentielle de certains gènes comme *phnA2a*, *bphA3*, *nahD*, *xylE*, *ahdA1d*, *bphK* ou encore le régulateur potentiel codé par Orf597 en fonction des sources de carbone et d'énergie disponibles. Une forte expression des gènes *ahdA1d, bphK, xylE* et *xylC* en présence des deux HAP est également observée, en accord avec des études précédentes (Kim et Zylstra, 1999; Romine *et al.*, 1999a; Stolz, 2009). En effet, la présence d'un mélange de plusieurs HAP engendrerait une plus forte expression des différents groupes de gènes et donc une minéralisation de ces composés plus efficace et plus rapide. De plus, d'autres gènes (comme *nahD* et l'Orf597) semblent exprimés uniquement en présence de phénanthrène, indiquant leur spécificité. Il est à noter que cette spécificité a été démontré pour le gène *nahD* (Kim *et al.*, 1997).

L'application de cette biopuce pour une étude environnementale permettra donc d'évaluer la présence des gènes ciblés au sein des écosystèmes étudiés. Les sondes

développées permettront également d'estimer la diversité génique présente dans l'environnement étudié. Finalement, une partie de ces travaux a été valorisée et publiée dans la revue BMC Bioinformatics (Terrat *et al.*, 2010) (publication disponible en Annexe 7).

Tableau 21 : Similarité des séquences du gène *phnA1a* isolées par PCR du site de type sol pollué par des HAP.

		Résultats par l'approche BLASTx					
Nom de la séquence	**Longueur (bases)**	**Identité (%)**	**Expected Value**	**Organisme**	**Enzyme**	**Nom du gène**	**Numéro d'accession**
Seq1	480	62	$2e^{-58}$	*Burkholderia* sp. DBT1	Sous-unité d'oxygénase alpha	*dbtAc*	AF380367.1
Seq2	480	99	$3e^{-90}$	*Polaromonas naphthalenivorans CJ2*	Sous-unité d'oxygénase alpha	*nagAc*	AAZ93388.1
Seq3	479	99	$3e^{-90}$	*Polaromonas naphthalenivorans CJ2*	Sous-unité d'oxygénase alpha	*nagAc*	AAZ93388.1
Seq4	479	99	$3e^{-90}$	*Polaromonas naphthalenivorans CJ2*	Sous-unité d'oxygénase alpha	*nagAc*	AAZ93388.1
Seq5	480	99	$3e^{-90}$	*Polaromonas naphthalenivorans CJ2*	Sous-unité d'oxygénase alpha	*nagAc*	AAZ93388.1
Seq6	480	99	$3e^{-90}$	*Polaromonas naphthalenivorans CJ2*	Sous-unité d'oxygénase alpha	*nagAc*	AAZ93388.1

Chapitre III : Etude de la diversité métabolique et phylogénétique de la communauté bactérienne d'un écosystème pollué par des HAP

1. Introduction

Dans un contexte de changements globaux où les problèmes de pollutions sont récurrents, il est primordial de connaître non seulement l'étendue de la diversité des microorganismes présents dans notre environnement, mais aussi de pouvoir identifier rapidement ceux disposant de capacités métaboliques pouvant contribuer à la restauration de l'écosystème perturbé. Ainsi, dans le cadre de ce travail, nous nous sommes intéressés à caractériser la diversité fonctionnelle de la communauté bactérienne issue d'un écosystème terrestre contaminé par différents types de polluants, tels des HAP et certains autres composés aromatiques (Tableau 12).

L'identification des gènes codant les processus biologiques impliqués dans les voies de dégradation de polluants aromatiques a été réalisée à l'aide de la biopuce ADN fonctionnelle précédemment décrite, et développée grâce à l'outil Metabolic Design.

2. Estimation de la diversité fonctionnelle

La biopuce ADN fonctionnelle exploratoire, précédemment validée sur la souche EPA505, a été utilisée pour étudier les capacités métaboliques de biodégradation de la communauté microbienne du site pollué. L'analyse des résultats, après hybridation de l'ADN total extrait de l'environnement pollué, a permis de mettre en évidence la présence de 31 des gènes ciblés. Cependant, neuf gènes n'ont pas été détectés au sein de cet écosystème avec les oligonucléotides déterminés : *phnA1a*, *nahE*, *bphA1e*, *xylX*, *bphK*, *xylC*, *xylE*, l'orf758 et *catR*. Etonnamment, aucune sonde spécifique du gène *phnA1a* ne donne de signal avec l'ADN extrait de l'écosystème pollué. Or, ce gène *phnA1a*, qui code pour la sous-unité α de la dioxygénase initiale, est indispensable à la dégradation des HAP chez une grande majorité des microorganismes. Afin de pouvoir expliquer ce résultat, nous avons choisi d'amplifier par PCR ce gène, au sein de l'ADN total isolé de l'environnement d'intérêt. Les produits PCR amplifiés et clonés ont ensuite été séquencés et analysés (Tableau 21). Les séquences

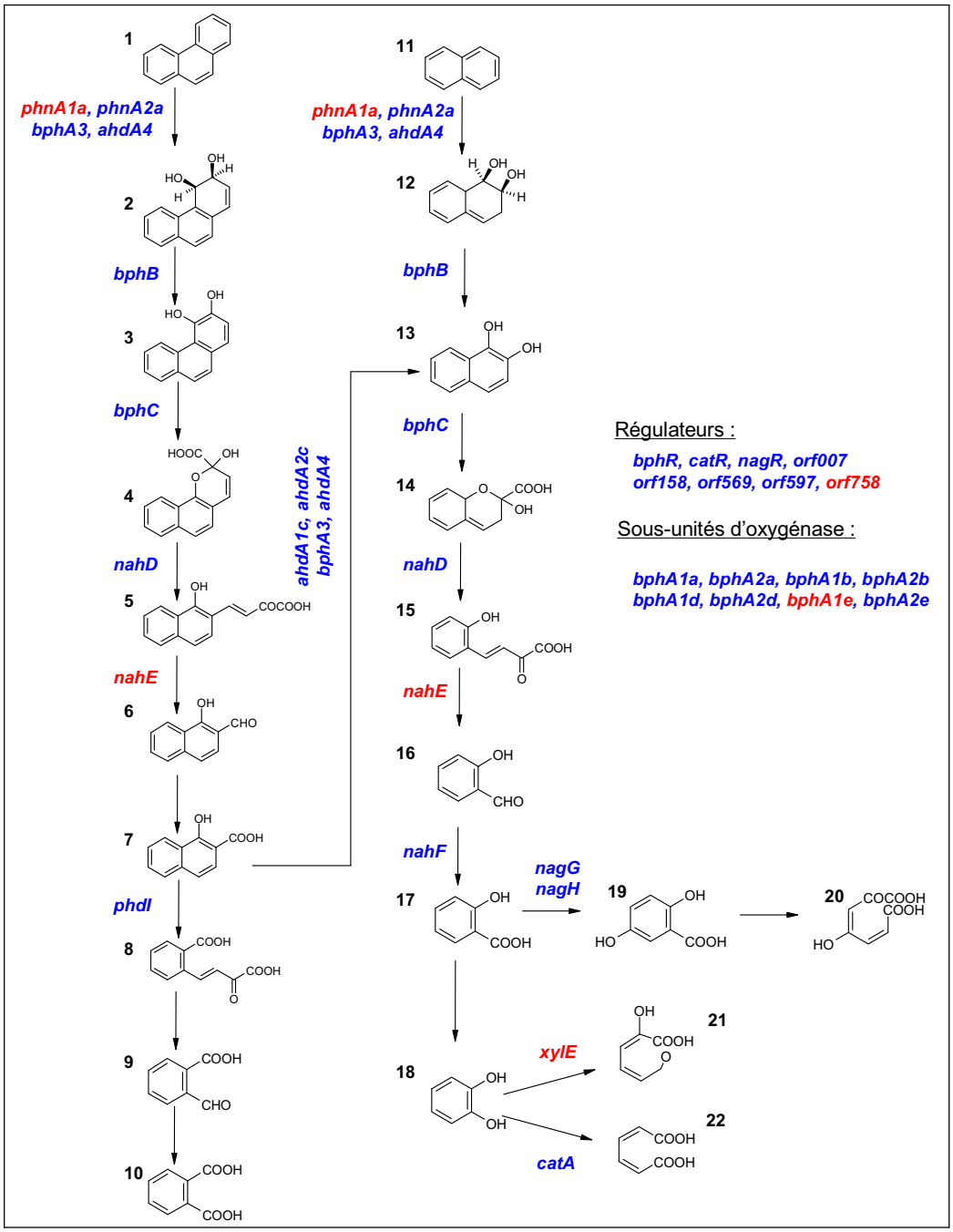


Figure 54 : Voies de dégradation potentielles détectées au sein de l'écosystème pollué à l'aide la biopuce ADN fonctionnelle.

Les flèches sans lettres peuvent représenter une ou plusieurs étapes enzymatiques et n'ont pas été ciblées. Les noms des gènes ciblés sont représentés sur le schéma, où les enzymes codées par ces derniers interviennent potentiellement. En bleu apparaissent les gènes détectés, en rouge, ceux non détectés. Sur le schéma sont également représentés les régulateurs ciblés et les sous-unités d'oxygénase ciblés. Certains gènes détectés d'autres voies de dégradation ne sont pas représentés.

Composés : (**1**) Phénanthrène ; (**2**) *cis*-3,4-phénanthrène dihydrodiol ; (**3**) 3,4-dihydroxyphénanthrène ; (**4**) 2-hydroxybenzo(*h*)chromène-2-carboxylate ; (**5**) 4-(1-hydroxynapht-2-yl)-2-oxobut-3-énoate ; (**6**) 1-hydroxy-2-naphtaldéhyde ; (**7**) 1-hydroxy-2-napthoate ; (**8**) *trans-o*-carboxybenzylidène pyruvate ; (**9**) 2-carboxybenzaldéhyde ; (**10**) ; (**11**) *o*-phthalate ; naphtalène ; (**12**) *cis*-1,2-naphtalène dihydrodiol ; (**13**) 1,2-dihydroxynaphtalène ; (**14**) 2-hydroxychromène-2-carboxylate; (**15**) *trans-o-*

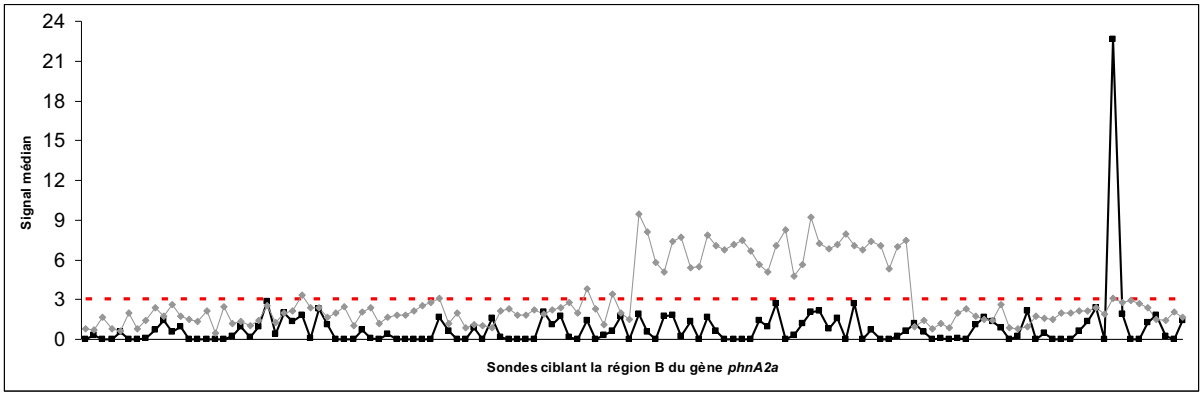
obtenues sont proches des gènes *nagAc* et *dbtAc*, codant pour des sous-unités α de dioxygénases initiales intervenant dans l'attaque initiale de polluants aromatiques. Cependant, ces séquences sont assez éloignées du gène *phnA1a* (car présentant moins de 70 % d'identité nucléique). L'analyse de la séquence de la région ciblée par les sondes déterminées avec Metabolic Design révèle l'existence de plusieurs mésappariements entre les séquences isolées et celle de la sonde, pouvant expliquer l'absence d'hybridation entre la sonde et les cibles. Ce résultat démontre donc la présence de dioxygénases initiales au sein de l'écosystème étudié, mais trop divergentes de celles ciblées avec notre approche pour être détectées.

Il est intéressant de noter que pour une grosse proportion des 31 gènes détectés, et pour une même région ciblée considérée, plusieurs sondes spécifiques ont donné des signaux au dessus du seuil de significativité. Ces résultats mettent en évidence la présence de plusieurs variants de ces gènes au sein de l'environnement étudié. Ces informations sur les capacités métaboliques endogènes peuvent être reliées à la présence de divers HAP détectés au sein de ce sol pollué. Ainsi, une forte présence dans le sol, de naphtalène (620 mg/kg de masse sèche), de phénanthrène (430 mg/kg de masse sèche), de fluoranthène (270 mg/kg de masse sèche) et de pyrène (210 mg/kg de masse sèche) ont été mises en évidence (Tableau 12). Si l'on prend l'exemple du phénanthrène, de nombreux gènes détectés codent pour des protéines impliquées dans sa dégradation (Figure 54). Ainsi, certaines bactéries au sein de l'environnement étudié semblent posséder la voie commune de dégradation, comme le montre la détection des gènes *phnA2a, bphA3, ahdA4, bphB, bphC* et *nahD* (bien que *phnA1a* et *nahE* ne soient pas mis en évidence avec les sondes déterminées). Pour ces gènes identifiés, de nombreux variants ont pu être détectés avec les différentes sondes déduites des sondes dégénérées. Par exemple, pour le gène *phnA2a*, sur les 128 sondes spécifiques ciblant la région B de ce gène, 37 montrent un signal supérieur au seuil défini, avec un signal maximum de 9,47 ± 0,70 (Figure 55 page suivante). Le contrôle utilisant les ARN extraits de la souche *Sphingomonas paucimobilis* sp. EPA505 montre bien l'hybridation spécifique avec la sonde s'appariant parfaitement avec le gène cible de la souche. De même, pour le gène *bphC*, 16 sondes donnent un signal positif (avec un SNR' maximum de 4,33 ± 1,14) parmi les 128 sondes spécifiques qui ciblent la région B de ce gène. Les microorganismes endogènes ont donc les capacités métaboliques pour dégrader en partie le phénanthrène et le naphtalène.

De plus, d'autres gènes ont été détectés, comme les gènes *phdI* (168 sondes positives sur 384)*, ahdA1c* (204 sur 1 024) et *ahdA2c* (18 sur 1 024). La caractérisation, du gène *phdI,* et des gènes *ahdA1c* et *ahdA2c,* respectivement spécifiques des voies de Kiyohara et de Evans (Figure 54), montre que le phénanthrène peut être dégradé par l'une ou l'autre des voies de

hydroxybenzylidène pyruvate ; (**16**) salicylaldéhyde ; (**17**) salicylate ; (**18**) catéchol ; (**19**) gentisate ; (**20**) maléylpyruvate ; (**21**) 2-hydroxymuconate semi aldéhyde ; (**22**) muconate.

Nom des gènes : *phnA1a* : codant une sous-unité α de la dioxygénase initiale putative ; *phnA2a* : codant une sous-unité β de la dioxygénase initiale putative ; *bphA3* : codant une sous-unité de ferrédoxine putative ; *ahdA4* : codant une ferrédoxine réductase putative ; *bphB* : codant une *cis*-dihydrodiol déshydrogénase putative ; *bphC* : codant une dihydroxynaphtalène dioxygénase putative ; *ahdA2c* : codant une petite sous-unité d'oxygénase putative ; *ahdA1c* : codant une grande sous-unité d'oxygénase putative ; *nahD* : codant une 2-hydroxychromène-2-carboxylate isomérase putative ; *nahE* : codant une dihydroxybenzylpyruvate aldolase putative ; *phdI* : 1-hydroxy-2-naphthoate dioxygénase ; *nahF* : Salicylaldéhyde déshydrogénase ; *nagG* : Grande sous-unité du salicylate 5-hydroxylase ; *nagH* : Petite sous-unité du salicylate 5-hydroxylase ; *xylE* : catéchol 2,3-dioxygénase ; *catA* : catéchol 1,2-dioxygénase.



<u>Figure 55 :</u> Intensité du signal d'hybridation obtenue pour chacune des 128 sondes ciblant la région B du gène *phnA2a*.

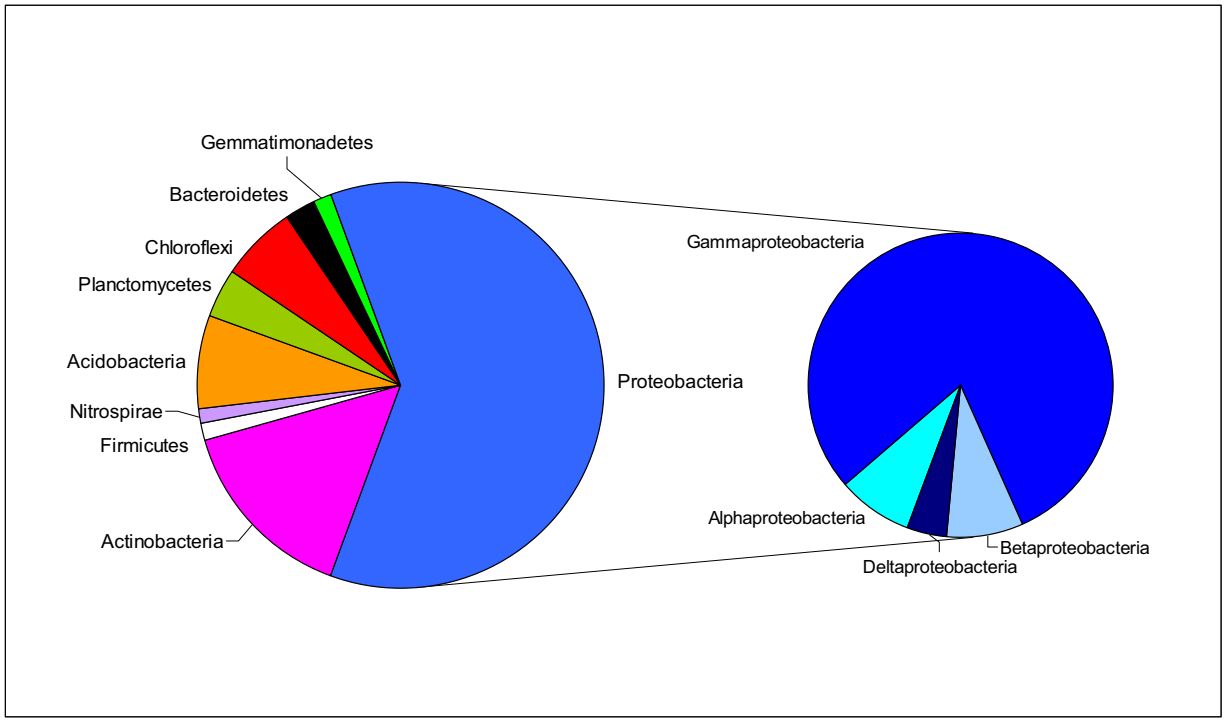
Les résultats ont été obtenus après hybridation : soit des ARN de la souche EPA505, extraits des cultures réalisées avec un mélange de phénanthrène et de fluoranthène (carrés noirs), soit de l'ADN total extrait de l'écosystème sol contaminé par des HAP (losanges gris). La ligne pointillée représente le seuil défini pour considérer le signal obtenu comme positif (SNR' > 3).

biotransformations, au sein de l'écosystème étudié. De plus, le fait que plusieurs sondes pour chacun de ces trois gènes donnent un signal positif suggère qu'il existe une communauté bactérienne diversifiée, capable d'assurer ces processus métaboliques. La voie de Evans (réalisée par les protéines codées par *ahdA1c, ahdA2c, bphA3* et *ahdA4*, tous détectés) permet de rejoindre la voie de dégradation du naphtalène. Cette voie nécessite d'autres enzymes spécifiques, comme NahF. Or, le gène codant cette enzyme a été mis en évidence avec l'approche biopuce fonctionnelle (Figure 54). Ainsi, avec 292 sondes donnant un signal supérieur au seuil fixé, une importante variabilité génique est vraisemblablement présente pour ce gène. La présence de *nahF*, mais aussi d'autres gènes intervenant dans la voie commune de dégradation nous montre que les communautés microbiennes présentes au sein du site d'intérêt peuvent dégrader le naphtalène, également fortement présent.

Ces voies aboutissent à la formation de composés devant rejoindre le métabolisme central pour permettre la production d'énergie. Les résultats d'hybridation révèlent la présence potentielle de trois gènes : *nagG* (avec 70 sondes positives sur 256 sondes totales), *nagH* (avec 17 sondes positives sur 128 sondes totales) et *nagR* (avec 13 sondes positives sur 384 sondes totales) codant des enzymes impliquées dans la voie spécifique de dégradation du gentisate (Figure 54). La caractérisation du gène *catA* (7 sondes donnant un signal positif sur 1 024 sondes totales) est en accord avec l'existence de la voie de clivage dite *ortho* également présente, comme le montre ces résultats (Figure 54). La troisième voie, celle dite voie de clivage *meta*, ne semble pas être présente au sein de cet environnement, comme le révèle les résultats du gène *xylE*, pour lequel aucune sonde ne donne un signal positif (Figure 54).

Les sous-unités d'oxygénases étant largement impliquées dans les voies de dégradation de composés aromatiques, et conférant aux différentes souches bactériennes la capacité de dégrader une large gamme de substrats, plusieurs couples ont également été détectés. L'analyse des résultats obtenus révèle une importante diversité de ces sous-unités d'oxygénase au sein du site étudié. Ainsi, 124 sondes répondent pour *bphA1a* (pour 512 sondes totales), 325 pour *bphA2a* (pour 512 sondes), 45 pour *bphA1b* (pour 128 sondes) et 4 pour *bphA2b* (pour 384 sondes), 258 sondes pour *ahdA1d* (pour 512 sondes) et 108 pour *ahdA2d* (pour 1 024 sondes) et enfin 372 sondes pour *bphA2e* (pour 1 024 sondes).

Enfin, certains régulateurs (répresseurs ou activateurs) potentiels ciblés sont détectés par les sondes déterminées. Ainsi, l'analyse des résultats obtenus montre que *bphR* (12 sondes donnant un signal positif sur 768 sondes totales), l'Orf007 (10 sondes sur 512 sondes totales), l'Orf158 (2 sondes sur 384 sondes totales), l'Orf569 (21 sondes sur 768 sondes totales) et l'Orf597 (65 sondes sur 256 sondes totales) ont été respectivement mis en évidence au sein du



<u>Figure 56 :</u> Affiliation taxonomique des séquences issues de la librairie de clones ADNr 16S obtenue à partir de l'environnement sol pollué par des HAP.

Les *Proteobacteria* représentent le phyla majoritaire (61,3% des séquences obtenues) suivies par les *Actinobacteria* (15%), les *Acidobacteria* (7,5%) et les *Chloroflexi* (6,3%). Le phyla des *Proteobacteria* est composé de *Gammaproteobacteria* (80%), de *Betaprotebacteria* (8%), d'*Alphaproteobacteria* (8%) et de *Deltaproteobacteria* (4%).

site pollué. Il reste cependant à déterminer précisément l'implication de ces différents régulateurs dans les voies de dégradation des HAP.

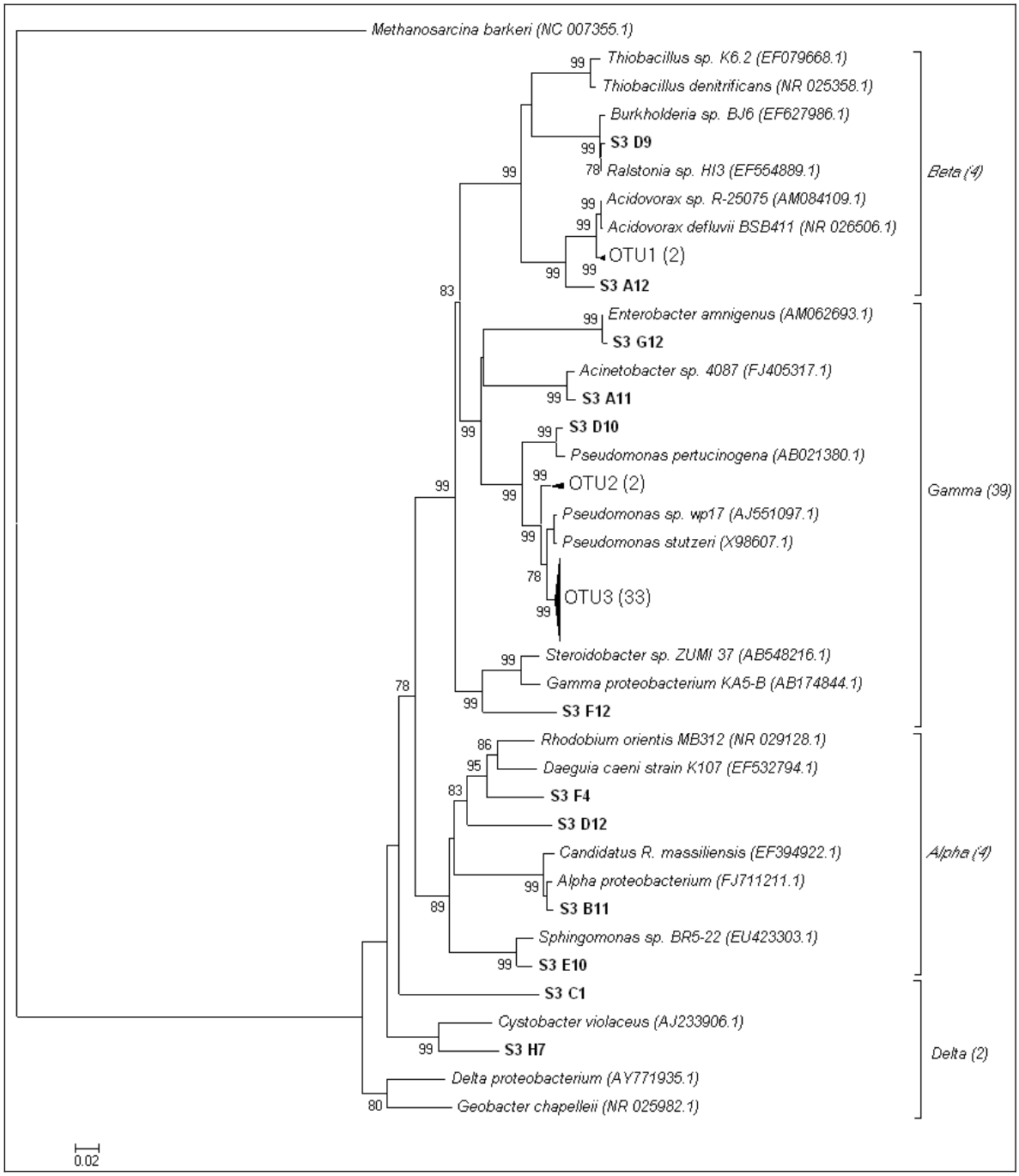
3. Conclusion

Les résultats acquis avec la biopuce fonctionnelle exploratoire ont permis de caractériser les potentialités métaboliques de dégradation de polluants aromatiques présentes au sein du site pollué, sans *a priori* sur les séquences, ou les organismes présents. Plusieurs voies de dégradation potentielles ont ainsi été détectées au sein de l'écosystème d'intérêt, comme celles du phénanthrène, de Kiyohara, du naphtalène, ou certaines voies basses (comme la voie du gentisate et celle de clivage dite *ortho*). Une grande diversité de sous-unités d'oxygénase a été également mise en évidence, ainsi que la présence de certains régulateurs potentiels des voies de dégradation des composés aromatiques. Avec la biopuce fonctionnelle exploratoire, il a donc été possible de reconstruire les capacités métaboliques globales de l'écosystème, mais aussi de montrer, par la caractérisation de plusieurs variants géniques pour chaque gène étudié, l'existence de différents groupes de bactéries possédant ces voies de biodégradation.

Il serait très intéressant de relier potentiellement les informations de structure des communautés microbiennes de ce sol pollué, à celles des potentialités métaboliques d'intérêt mises en évidence par la biopuce ADN fonctionnelle. C'est pourquoi une étude préliminaire de la structure de ces communautés a été réalisée par deux approches différentes : une approche de clonage-séquençage du gène codant l'ARNr 16S, et une approche par biopuce ADN taxonomique.

4. Etudes préliminaires et perspectives

La construction d'une banque de clones d'ADNr 16S provenant du sol pollué et son séquençage a été initiée et a permis de déterminer une fraction de la communauté bactérienne de cet environnement. 80 séquences ont pour le moment été analysées. Ces séquences ont été subdivisées en 45 Operational Taxonomic Units (OTU), chaque OTU regroupant toutes les séquences présentant plus de 97% de similarité (Figure 56). Bien évidemment l'analyse de ce nombre limité de clones ne permet pas de décrire l'entière diversité bactérienne. Les indices ACE et Chao1 montrent respectivement que le nombre réel d'OTUs avoisine 345 et 325.



<u>Figure 57 :</u> Arbre phylogénétique des séquences d'ADNr 16S montrant l'emplacement des OTUs de la communauté bactérienne issue de l'environnement pollué par des HAP au sein du phylum des *Proteobacteria*.

Les séquences ont été alignées à l'aide du logiciel ClustalW intégré au package MEGA 4.0 (Tamura *et al.*, 2007). Les OTUs définis sont représentés en gras et les nombres entre parenthèses indiquent le nombre de séquences appartenant à chaque OTU. L'arbre a été construit par la méthode du plus proche voisin (Saitou et Nei, 1987), et est enraciné à l'aide de la séquence du gène codant pour l'ARNr 16S de *Methanosarcina barkeri*. Sa robustesse a été évaluée par un bootstrap effectué sur 1 000 réplicats et seules les valeurs supérieures à 75 sont représentées. Les séquences de références sont suivies par le numéro d'accession GenBank entre parenthèse. La barre d'échelle représente le nombre de substitution pour 100 sites par unité de longueur de branche.

Cependant, l'analyse phylogénétique préliminaire des séquences révèle que le phylum le plus abondant est celui des *Proteobacteria* (61,3 %), suivi des *Actinobacteria* (15 %), des *Acidobacteria* (7,5 %) et des *Chloroflexi* (6,3 %) (Figure 56). Les phyla *Bacteroidetes*, *Firmicutes*, *Gemmatimonadetes*, *Nitrospirae* et *Planctomycetes* représentent chacun moins de 5 % des séquences affiliées. Parmi les *Gammaproteobacteria,* qui constituent la classe bactérienne majoritaire au sein des *Proteobacteria* (48,8 % de la communauté totale), un OTU est fortement représenté avec 33 clones. Ce dernier, est affilié aux bactéries appartenant au genre *Pseudomonas* (97 et 99 % d'identité avec les séquences de ce genre) (Figure 57). Ce dernier, à l'instar de l'OTU 2, est affilié aux bactéries appartenant au genre *Pseudomonas* (97 et 99 % d'identité avec les séquences de ce genre) (Figure 57). L'arbre phylogénétique obtenu montre également que, parmi les *Betaproteobacteria*, les OTUs 1 et S3_D9 ont les espèces relatives les plus proches qui appartiennent respectivement aux genres *Acidovorax* (99 % d'identité avec la séquence *Acidovorax* sp. R-25075 (AM084109)) et *Ralstonia* (99 % d'identité avec la séquence de *Ralstonia* sp. HI3 (EF554889)).

En parallèle, l'hybridation de l'ADN génomique total sur la biopuce taxonomique a révélé, au sein de l'environnement contaminé par des HAP, la présence de 45 genres bactériens appartenant à 5 phyla différents, à savoir : les *Proteobacteria*, les *Actinobacteria*, les *Bacteroidetes*, les *Firmicutes*, et les *Spirochaetes* (Tableau 22 page suivante). Les résultats de la biopuce taxonomique montrent que 25 genres bactériens ont été détectés au sein du phylum des *Proteobacteria* (Tableau 22). Plus précisément, parmi les *Gammaproteobacteria,* classe bactérienne majoritaire en terme de nombre de genres détectés avec la biopuce, 17 genres non identifiés par l'approche de clonage-séquençage du gène codant l'ARNr16S ont été retrouvés. Parmi ceux-ci, les genres comme *Marinospirillum* ou *Endozoicomonas*, ont été, jusqu'à présent, identifiés au sein d'écosystèmes aquatiques. Les résultats ont de plus permis de montrer l'existence de bactéries appartenant aux genres *Aeromonas, Halomonas* et *Pseudoalteromonas*, plus souvent retrouvés au sein de sols pollués. Deux autres classes de *Proteobacteria* ont également été détectées avec la biopuce taxonomique : la classe des *Alphaproteobacteria* (avec les genres *Martelella, Pseudaminobacter, Rhodopseudomonas* et *Roseovarius*), et celle des *Betaproteobacteria* (avec les genres *Janthinobacterium, Herbaspirillum, Uruburuella* et *Vogesella*).

Pour le phylum des *Actinobacteria*, 15 genres bactériens différents sont détectés, faisant de lui le second phylum le plus important en termes de diversité (Tableau 22). Parmi ceux-ci, les genres *Microterricola*, *Mycobacterium*, *Gordania,* ou encore *Rhodococcus* sont retrouvés, et sont connus pour être fortement représentés au sein des écosystèmes sols.

Tableau 22 : Genres bactériens détectés avec la biopuce ADN taxonomique au sein de l'environnement pollué par des HAP.

Phyla bactériens	Classes bactériennes	Genres détectés avec la biopuce taxonomique
Proteobacteria	***Alphaproteobacteria***	*Martelella, Pseudaminobacter, Rhodopseudomonas, Roseovarius*
	Betaproteobacteria	*Janthinobacterium, Herbaspirillum, Uruburuella, Vogesella*
	Gammaproteobacteria	*Aeromonas, Amphritea, Azotobacter, Azorhizophilus, Bermanella, Endozoicomonas, Erwinia, Halomonas, Haererehalobacter, Kluyvera, Legionella, Marinospirillum, Pantoea, Pectobacterium, Proteus, Pseudoalteromonas, Zobellella*
Actinobacteria	***Actinobacteria***	*Amycolatopsis, Actinomyces, Corynebacterium, Gordonia, Herbidospora, Humibacillus, Humicoccus, Krasilnikovia, Mycobacterium, Microbispora, Microterricola, Microtetraspora, Rhodococcus, Schumannella, Streptoalloteichus*
Bacteroidetes	***Bacteroidia***	*Bacteroides*
	Sphingobacteria	*Cytophaga*
Firmicutes	***Bacilli***	*Brevibacillus*
	Clostridia	*Lactonifactor*
Spirochaetes	***Spirochaetes***	*Treponema*

Enfin, les phyla *Bacteroidetes* (avec les genres *Bacteroides* et *Cytophaga*), *Firmicutes* (avec les genres *Brevibacillus* et *Lactonifactor*) et *Spirochaetes* avec le genre *Treponema*, ont également été identifiés au sein de l'écosystème sol contaminé par des polluants aromatiques à l'aide de l'approche biopuce ADN.

Ces données préliminaires nous permettent déjà d'émettre certaines hypothèses sur le fonctionnement des communautés au sein du site d'intérêt. Cependant, la confrontation des résultats complets sur la diversité phylogénétique et les potentialités métaboliques pourra au final confirmer ou non ces hypothèses. Ces données pourront par la suite orienter des optimisations pour les processus de biodégradation au sein de cet écosystème pollué.

DISCUSSION

DISCUSSION

L'écologie microbienne s'intéresse au fonctionnement des écosystèmes dans leur globalité, ceci afin d'acquérir de nouvelles connaissances, notamment sur les interactions entre les microorganismes. Pour atteindre cet objectif, les approches de biologie moléculaire se sont imposées car, tout en permettant de s'affranchir des biais liés aux méthodes culturales, elles assurent une vision rapide et globale de la diversité. Parmi elles, on retrouve la technique des biopuces ADN, fonctionnelles ou phylogénétiques, dont l'utilisation tend à se généraliser de plus en plus, pour étudier le fonctionnement d'écosystèmes complexes (Liang *et al.*, 2009a; Liang *et al.*, 2009b; Sei *et al.*, 2009; Iwai *et al.*, 2010; Rastogi *et al.*, 2010).

Toutefois, ces biopuces présentent encore certaines limites, liées notamment à la nature des sondes, qui ne permettent de cibler que ce qui a déjà été identifié au niveau génique. Le défi actuel est donc d'essayer de définir des sondes présentant un caractère exploratoire, c'est-à-dire capables d'appréhender toute la diversité génique présente au sein de l'écosystème étudié, même si celle-ci est encore inconnue. L'objectif principal de cette étude a donc été le développement d'une biopuce ADN fonctionnelle, permettant d'étudier les capacités métaboliques d'un écosystème pollué, et ce, sans connaissance préalable des séquences des gènes ciblés. Pour cela, une nouvelle approche a été développée et implémentée, dans un outil informatique, appelé Metabolic Design.

1. Reconstruction métabolique et fouille de données

Préalablement à la détermination des sondes, l'outil Metabolic Design a été développé pour reconstruire et visualiser un processus biologique à façon. Cette reconstruction permet ensuite d'effectuer, pour chacune des étapes du processus considéré, une fouille de données afin de rechercher l'ensemble des séquences protéiques pouvant potentiellement assurer cette étape biologique. Pour effectuer cette fouille de données, le choix de la séquence de référence est donc crucial. En effet, il est indispensable de s'appuyer sur des données validées biologiquement (et/ou appuyées par une bibliographie précise), pour choisir les protéines de référence. Par défaut, le logiciel Metabolic Design utilise donc les séquences répertoriées dans la base de données Swiss-Prot. Il est également possible d'utiliser des données personnelles, non encore disponibles au sein des bases. Ces séquences protéiques de référence sont ensuite utilisées comme séquences requêtes pour effectuer une fouille de données contre différentes banques, qu'elles soient publiques (comme Swiss-Prot et/ou TrEMBL), ou privées. Cette

recherche de similarité permet de s'affranchir des erreurs d'annotations fonctionnelles fréquemment rencontrées pour les séquences des bases de données généralistes. Donc, avec l'outil Metabolic Design, l'utilisateur a la possibilité de reconstruire ses propres processus biologiques, en combinant des informations de plusieurs sources : banques de données publiques, privées et/ou des informations bibliographiques.

Une telle flexibilité est un des principaux avantages de cette approche, en comparaison des outils de reconstruction métabolique, car ils utilisent la plupart du temps des bases statiques. Néanmoins, certains outils comme PathwayVoyager (Altermann et Klaenhammer, 2005), KGML-ED (Klukas et Schreiber, 2007), KEGGanim (Adler *et al.*, 2008), FMM (Chou *et al.*, 2009) et MinPath (Ye et Doak, 2009), permettent également une reconstruction graphique des voies métaboliques en intégrant des données personnelles. Cependant, pour effectuer la fouille de données, ces logiciels utilisent la base de données KEGG, et ne laissent pas la possibilité d'utiliser d'autres banques de données (Kanehisa *et al.*, 2008). Cette base KEGG, bien qu'étant l'une des plus complètes actuellement, ne regroupe que les données provenant d'organismes identifiés et caractérisés. Si l'on prend l'exemple des voies de dégradation des HAP, les données disponibles dans cette banque sont peu nombreuses. Ainsi, une seule voie incomplète, décrivant la dégradation du phénanthrène, du naphtalène et de l'anthracène est disponible dans KEGG-PATHWAY. De plus, les données de séquences issues d'environnements, et non encore annotées, ne sont pas intégrées, rendant la fouille de données et la reconstruction réalisées avec les logiciels cités précédemment non exhaustive.

La fouille de données réalisée par Metabolic Design se base uniquement sur la recherche de similarité entre la séquence de la protéine de référence, et celles des bases utilisées. Cependant, par cette approche, il est difficile de pouvoir identifier les protéines ayant une même fonction biologique, tout en présentant des séquences fortement divergentes (Singh, 2010). Une amélioration de Metabolic Design serait donc d'optimiser l'étape de fouille de données. Plusieurs possibilités sont ainsi envisagées :

(1) Une première solution serait de modifier les paramètres de l'étape de BLASTp, en utilisant par exemple, des matrices plus adaptées telle la matrice BLOSUM45, ou encore en optimisant certains paramètres comme la diminution des valeurs des pénalités d'ouvertures et d'extensions de gaps accordées lors de l'alignement. De telles modifications amélioreraient la fouille de données, pour ainsi identifier des séquences fortement divergentes, mais ayant une même fonction biologique.

(2) Il serait, également intéressant de s'appuyer sur l'algorithme PSI-BLAST (ou Position Specific Iterative BLAST), en remplacement du BLASTp. Cet algorithme s'appuie sur

l'utilisation d'un alignement multiple de séquences (issues d'une première itération de BLASTp), afin de générer un profil de conservation des séquences (un score élevé est donné à une région fortement conservée, et un score de 0 à une région faiblement conservée). Ce profil est ensuite utilisé pour réaliser de nouvelles itérations de BLASTp, ce qui permet d'identifier des séquences divergentes d'une même famille protéique, non détectées par une fouille de données utilisant l'approche classique de BLASTp.

L'application de ces approches à la détermination de nouvelles sondes pour *phnA1a* permettrait potentiellement d'étendre l'identification de séquences divergentes de cette famille enzymatique, et donc d'assurer l'identification des séquences caractérisées par l'approche PCR.

Enfin, prendre en compte au sein du logiciel Metabolic Design l'identification de séquences orthologue améliorerait la qualité de la fouille de données effectuée. En effet, l'ajout de modules complémentaires comme Ortholuge, ou Ortho-MCL (Chen *et al.*, 2006; Fulton *et al.*, 2006), assurerait une meilleure identification de protéines divergentes ayant une même fonction biologique parmi un jeu de données très important et difficile à trier manuellement. Cette optimisation serait ainsi particulièrement utile dans la cas de familles enzymatiques complexes, comme celles des dioxygénases, largement impliquées dans la dégradation de composés aromatiques comme les HAP (Kweon *et al.*, 2008).

2. Détermination de sondes exploratoires

La fouille de données constitue le point de départ pour la détermination de sondes exploratoires, qui peut être réalisée pour tout type de gène codant une protéine. L'originalité de cette détermination de sondes est de s'appuyer sur un alignement protéique. En effet, en se basant sur ce type d'alignement, il est possible de considérer pour chaque site moléculaire les différents acides aminés pouvant être rencontrés, et ensuite, de définir la séquence de la sonde par traduction inverse, en prenant en compte la dégénérescence complète du code génétique. Il est alors possible d'explorer toute la diversité génique potentielle, en définissant des sondes ciblant à la fois des séquences connues, et des séquences encore inconnues, codant toutes pour la même séquence protéique. Cet aspect exploratoire n'est pas intégré par les outils permettant de développer des sondes pour biopuces ADN fonctionnelles, comme OligoArray 2.0, YODA ou HPD (Lemoine *et al.*, 2009). Etonnamment, cette approche est pourtant banalisée, pour la détermination d'amorces utilisées dans les approches de PCR en écologie microbienne (Jabado *et al.*, 2006). L'outil Metabolic Design intègre également des résidus Inosines dans la

séquence des sondes dégénérées. En effet, ces bases modifiées peuvent être utilisées dans la fabrication de sondes des biopuces de type *ex situ*, permettant ainsi de diminuer le nombre de sondes différentes déposé sur le support solide au sein d'un même spot.

La sélection des sondes avec Metabolic Design n'est cependant basée que sur certains critères définis par l'utilisateur (taille, dégénérescence, composition en Inosines). La prise en compte d'autres paramètres (comme le Tm, certains paramètres thermodynamiques, la nature de la séquence, …) qui jouent un rôle prepondérant dans l'efficacité des sondes permettrait d'accroître leur sensibilité et leur homogénéité. Ainsi, la sélection des sondes en fonction des résultats du calcul du Tm, permettrait d'harmoniser leurs réponses durant l'hybridation (Lemoine *et al.*, 2009). En effet, il est primordial que les Tm de l'ensemble des sondes soient compris dans une gamme relativement réduite. Il serait également essentiel d'évaluer la stabilité de structures secondaires potentielles (duplex sonde/cible, structures tige-boucle au sein de la sonde, formation d'homo dimères), en calculant certains paramètres thermodynamiques, comme leur énergie libre de Gibbs respective (Pozhitkov *et al.*, 2006; Pozhitkov *et al.*, 2007; Li *et al.*, 2008; Arslan et Laurenzi, 2009; Mueckstein *et al.*, 2010). Ces calculs permettraient de définir potentiellement la plus stable de ces structures et faciliterait le choix des sondes (Pozhitkov *et al.*, 2007; Lemoine *et al.*, 2009). En effet, la faible sensibilité des sondes ciblant le gène *ahdA4* pourrait potentiellement s'expliquer par les valeurs de ces paramètres, non pris en compte dans la sélection des sondes. Pourtant, il est encore difficile d'évaluer ces critères de manière précise, surtout dans des conditions d'hybridations sonde-cible sur un support solide (Pozhitkov *et al.*, 2007). De même, la nature de la séquence (composition en bases et succession de bases) (Royce *et al.*, 2007) peut également jouer sur la sensibilité des sondes. Tous ces paramètres devraient donc être considérés pour encore améliorer la qualité et l'homogénéité des jeux de sondes exploratoires déterminés avec Metabolic Design.

Après avoir défini un jeu de sondes, l'outil Metabolic Design va en évaluer la spécificité, par rapport à des séquences potentiellement présentes au sein des environnements étudiés. Cette étape est réalisée par la comparaison de la séquence des sondes définies, avec une base de données répertoriant l'ensemble des séquences des divisions procaryotes (PRO), champignons (FUN) et environnement (ENV) de la base EMBL. Les biopuces devant permettre d'identifier les gènes d'intérêt, mais également d'étudier la modulation de leur expression au cours d'un processus biologique, les hybridations s'effectuent donc avec des cibles correspondant uniquement aux ARNm. Dans ce cas, les tests de spécificité ne doivent prendre en compte que les séquences potentiellement transcrites. C'est dans ce but que la

Base de Données Spe a été développée. En effet, celle-ci ne contient que les CDS et les régions UTR 5' et 3' des divisions précédemment citées.

D'autres outils de détermination de sondes effectuent des tests de spécificité en recherchant des séquences pouvant éventuellement provoquer des hybridations croisées. Ces tests s'appuient la plupart du temps sur des bases simples comme le génome d'une ou plusieurs souches, mais jamais sur des données aussi exhaustives que celles présentes dans la Base de Données Spe (Lemoine *et al.*, 2009). De plus, cette banque de données, comme les autres bases utilisées dans Metabolic Design, sont évolutives, et peuvent aussi être remplacées par des bases plus adaptées aux études réalisées. Par exemple, il est possible de remplacer la Base de Données Spe par une base de données issue d'un séquençage de métagénomique obtenue sur un écosystème particulier (Stenuit *et al.*, 2008). Pour évaluer la spécificité d'une sonde, Metabolic Design, comme de nombreux autres outils, utilise l'algorithme de recherche BLAST. Il serait néanmoins nécessaire d'optimiser ce test, en prenant par exemple en considération la nouvelle matrice optimisée pour l'étape de BLASTn décrite récemment (Eklund *et al.*, 2009), permettant d'assurer une identification plus fiable des hybridations croisées potentielles. Ainsi, une simple diminution du score alloué à la conservation des bases adénine et thymine (A-A et T-T, avec un nouveau score de 2, à la place d'un score de 5), permet une meilleure prédiction des hybridations croisées potentielles. En effet, ces deux bases contribuent moins, d'un point de vue énergétique, à l'affinité de la formation du duplex sonde/cible, que les bases guanine et cytosine.

3. Conception et validation d'une biopuce ADN fonctionnelle exploratoire

Metabolic Design a été utilisé pour définir des sondes ciblant un total de 40 protéines (39 étant impliquées dans les voies métaboliques de dégradation de composés aromatiques, le dernier étant le gène *gyrB* qui est un gène de ménage), en se basant principalement sur les informations de génomique disponibles pour les genres *Sphingomonas, Pseudomonas, Nocardioides* et *Ralstonia* (Habe et Omori, 2003; Pinyakong *et al.*, 2003a; Vandecasteele, 2005; Stolz, 2009). Une biopuce a ainsi été élaborée, composée de 39 216 sondes spécifiques, déduites de 72 sondes dégénérées. La souche bactérienne, *Sphingomonas paucimobilis* sp. EPA505, connue pour dégrader divers HAP, a ensuite été utilisée afin de valider la pertinence de l'approche, c'est-à-dire pour démontrer la sensibilité et la spécificité des sondes.

La validation s'est faite par l'étude de huit gènes (*phnA1a, phnA2a, ahdA1c, ahdA2c, ahdA4, bphB, bphA3* et *bphC*) qui codent pour des enzymes de dégradation de HAP, potentiellement présentes et exprimées en présence de HAP, par la souche modèle. Les résultats montrent qu'à l'exception des sondes ciblant le gène *ahdA4*, celles-ci permettent de mettre en évidence l'expression des gènes lorsque la souche est mise en culture en présence d'un mélange de phénanthrène et de fluoranthène. Des résultats similaires ont également été obtenus après hybridation des ARN extraits de cultures réalisées en présence d'un seul des deux HAP.

Il faut également noter que pour les sept gènes dont l'expression a pu être détectée, les différentes sondes ciblant plusieurs régions géniques ne donnent pas toutes des signaux au dessus du seuil de détection. C'est par exemple le cas des sondes ciblant la région A du gène *phnA2a*, où aucune sonde spécifique ne donne un signal, alors qu'une sonde spécifique ciblant la région B donne un signal important. De tels résultats confortent donc la nécessité de prendre en compte les contraintes thermodynamiques lors de l'élaboration des sondes et/ou de réaliser des validations préalables (Pozhitkov *et al.*, 2006; Pozhitkov *et al.*, 2007; Royce *et al.*, 2007; Li *et al.*, 2008; Arslan et Laurenzi, 2009; Mueckstein *et al.*, 2010). Nous avons cependant démontré dans la majorité des cas que la détermination des sondes était très spécifique. En effet, parmi l'ensemble des sondes déduites de chaque sonde dégénérée, seul un faible nombre de sondes donne un signal significatif, témoignant de la formation des duplex sonde/cible. Ayant pris en compte toute la dégénérescence possible de la région étudiée, il a pu être montré que ces sondes spécifiques possèdent, dans la plupart des cas, la séquence rigoureusement complémentaire aux séquences ciblées.

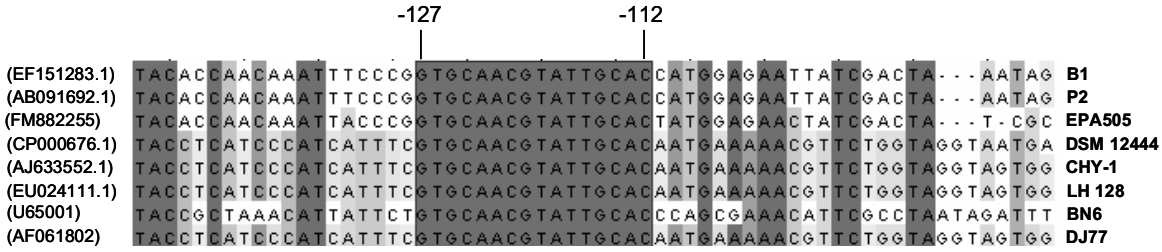
Ces résultats, tout en confirmant la spécificité des oligonucléotides développés, appuient donc la pertinence de notre stratégie pour déterminer des sondes exploratoires. En effet, celles-ci assurent l'identification de gènes, sans *a priori* sur leurs séquences, ce qui est un des aspects cruciaux de la détermination de sondes pour biopuce ADN, appliquée à l'étude d'environnements complexes (Kane *et al.*, 2000; Rhee *et al.*, 2004; Rimour *et al.*, 2005; Gentry *et al.*, 2006; Liebich *et al.*, 2006; Pozhitkov *et al.*, 2007).

Il était également important de s'assurer que les sondes développées avec l'outil Metabolic Design présentaient une sensibilité suffisante pour réaliser un suivi semi-quantitatif de l'expression des gènes. Pour cela, une étude de l'expression des gènes ciblés a été réalisée au cours de différents processus de biodégradation, par deux approches en parallèle. Les résultats obtenus avec les biopuces corroborent ceux de l'approche de PCR quantitative. En effet, il a été possible de mettre en évidence, par les deux approches, le profil d'expression des

gènes ciblés, en réponse à la présence de HAP, et de caractériser des modulations différentes selon la source carbonée. Ces résultats confirment donc que les sondes développées sont, dans la plupart des cas, suffisamment sensibles pour suivre l'expression de gènes d'intérêt, aspect indispensable pour des applications en écologie microbienne (Rhee *et al.*, 2004; Liang *et al.*, 2009a; Liang *et al.*, 2009b). De plus, il serait possible d'optimiser la sensibilité et la spécificité des sondes développées, en utilisant la stratégie « GoArrays » développée au sein de l'équipe (Rimour *et al.*, 2005). Cette approche consiste en la concaténation de deux sondes courtes (par un linker aléatoire de quelques bases), ciblant deux régions séparées d'un même gène, et ce pour déterminer une sonde longue composite, alliant à la fois l'aspect sensibilité des sondes longues et l'aspect spécificité des sondes courtes.

4. Caractérisation des capacités métaboliques et des régulations géniques chez la souche *Sphingomonas paucimobilis* sp. EPA505

Tout en confortant l'intérêt de l'approche Metabolic Design, la validation a aussi permis d'accroître nos connaissances sur les potentialités de la souche EPA505 à métaboliser différents HAP. Cette étude a notamment mis en lumière une métabolisation différentielle des substrats disponibles. En effet, en présence des deux HAP utilisés, la souche consomme préférentiellement le phénanthrène, par rapport au fluoranthène. La complexité des composés pourrait expliquer cette consommation. En effet, le phénanthrène, molécule plus simple (composée de trois cycles) que le fluoranthène (composée de quatre cycles), est dégradé préférentiellement, comme plusieurs études l'on déjà démontré (Dimitriou-Christidis et Autenrieth, 2007; Desai *et al.*, 2008). Une autre hypothèse est liée à l'affinité de l'enzyme pour les deux substrats (Prak et Pritchard, 2002). Ainsi, les deux HAP entrant en compétition, le phénanthrène est favorisé, car plus spécifique. Cette consommation préférentielle pourrait donc être reliée au spectre de substrat de l'enzyme multimérique impliquée dans l'attaque initiale de ces deux HAP. Ainsi, plusieurs autres dioxygénases ont été identifiées chez différentes souches de *Sphingomonas*, et leurs spectres de substrats identifiés. Par exemple, l'enzyme PhnI de la souche *Sphingomonas* sp. CHY-1 peut dégrader le naphtalène, le phénanthrène, l'anthracène et le chrysène, mais pas le fluoranthène (Demaneche *et al.*, 2004). De plus, la structure cristallographique de PhnI a récemment été obtenue et analysée pour la souche CHY-1 (Jakoncic *et al.*, 2007b). Les résultats cristallographiques révèlent la présence de deux boucles, L1 et L2, impliquées dans l'activité catalytique de l'enzyme PhnI, et plus précisément dans la reconnaissance du substrat. Les acides aminés les plus importants au sein



<u>Figure 58 :</u> Caractérisation de séquences régulatrices par alignement multiple des régions intergéniques des gènes *xylX* et *bphC* de plusieurs espèces bactériennes.
Les séquences proviennent des souches *Sphingobium yanoikuyae* B1 (EF151283), *Sphingobium* sp. P2 (AB091692), *Sphingomonas paucimobilis* sp. EPA505 (FM882255), *Novosphingobium aromaticivorans* DSM 12444 (CP000676), *Sphingomonas* sp. CHY-1 (AJ633552), *Sphingomonas* sp. LH128 (EU024111), *Sphingomonas xenophaga* BN6 (U65001) et *Pseudomonas* sp. DJ77 (aussi appelé *Sphingobium chungbukense*) (AF061802). Les positions sont données par rapport à la position +1 du gène *bphC*.

de ces boucles sont : Leu223, Leu226, Ile253 et Ile260. Et, si l'on compare les séquences protéiques de la souche CHY-1 avec la souche EPA505, les acides aminés Leu223 et Ile253 (pour CHY-1) ont été remplacés par Phe223 et Val253 (pour EPA505). Il est important de noter que la phénylalanine peut jouer un rôle dans les interactions hydrophobes (de par son cycle aromatique), ce qui pourrait expliquer les différences de spectre de substrat entre ces deux souches. Cette hypothèse reste cependant à vérifier par des analyses plus précises, en utilisant notamment des approches de mutagenèse sur ces résidus particuliers.

L'étude de l'expression des gènes ciblés montre également qu'en fonction de la source carbonée, ceux-ci sont différentiellement exprimés. Cette régulation fine de l'expression des gènes codant des enzymes impliquées dans la dégradation des HAP est encore méconnue et peu décrite (Pinyakong *et al.*, 2003a; Stolz, 2009). Par analyse des séquences obtenues, une région conservée potentiellement impliquée dans ces régulations a néanmoins été mise en évidence. La comparaison de la région intergénique de 289pb, située entre les gènes *xylX* et *bphC*, (Figure 49, A), avec plusieurs autres séquences, déjà identifiées, des genres *Sphingomonas* et *Pseudomonas*, montre une forte conservation de cette région (Figure 58). Cette région semble de plus, constituer une séquence palindromique (**GTGCAA**nnnn**TTGCAC**) qui pourrait intervenir dans la liaison de diverses protéines régulatrices, comme par exemple BphR, de la famille des régulateurs NtrC, ou encore des régulateurs appartenant aux familles LysR (généralement des activateurs) ou MucR (généralement des répresseurs) (Rothmel *et al.*, 1990; Romine *et al.*, 1999a). En effet, ces régulateurs agissent pour la plupart sous formes d'homodimères, en se liant à des séquences palindromiques (Huffman et Brennan, 2002). Ils sont donc vraisemblablement mis en jeu dans la régulation fine de l'expression de ces gènes codant des enzymes impliquées dans la dégradation des HAP (Stolz, 2009).

L'importance de la nature du substrat sur l'expression des gènes a aussi été démontrée pour certains des autres gènes ciblés avec la biopuce fonctionnelle exploratoire. En effet, le gène *nahD* (codant pour 2-hydroxychromène-2-carboxylate isomérase) est spécifiquement exprimé en réponse à la présence de phénanthrène, ou d'un mélange des deux HAP, mais pas en présence de fluoranthène. Il a été démontré par inactivation du gène *nahD*, chez *Sphingomonas yanoikuyae* B1 (proche de la souche EPA505), que ce dernier code une protéine indispensable à la voie de dégradation du naphtalène (Kim *et al.*, 1997). Or, la dégradation du phénanthrène chez plusieurs espèces de *Sphingomonas* conduit à la formation de 1,2-dihydroxynaphtalène, rejoignant ainsi la voie de dégradation du naphtalène (Pinyakong *et al.*, 2003a). Les résultats d'expression de ce gène chez EPA505 sont donc en accord avec

les données acquises chez la souche B1, et l'enzyme codée par *nahD* chez notre souche modèle ne serait donc pas impliquée dans la dégradation du fluoranthène, ce dernier étant probablement dégradé par une autre voie métabolique (Singleton *et al.*, 2009).

De même, certains couples de sous-unités d'oxygénase (comme *ahdA1d-ahdA2d*) sont exprimés en réponse à une exposition au mélange de HAP. Leur expression avait déjà été démontrée en présence de phénanthrène seul chez la souche *Sphingomonas* sp. P2 (Pinyakong *et al.*, 2003b). C'est également le cas d'autres gènes, comme *bphK, xylC* ou *xylE*. Par exemple, les résultats d'expression du gène *xylE*, codant une catéchol-2,3-dioxygénase intervenant dans la voie de clivage dite *meta* (SNR' = 6,77 ± 1,73; SNR' = 5,62 ± 0,44 et SNR' = 39,96 ± 4,95 respectivement obtenus avec les ARN extraits en présence de phénanthrène, de fluoranthène et d'un mélange de ces deux composés) révèlent une forte expression en présence d'un mélange des deux HAP. Ces résultats appuient l'hypothèse selon laquelle la présence d'un mélange de composés aromatiques permet une expression d'un plus grand nombre de clusters de gènes (et potentiellement d'une expression plus élevée), et donc une minéralisation des composés plus efficace et plus rapide grâce aux différentes isoenzymes produites (Romine *et al.*, 1999a; Stolz, 2009). Cette particularité serait probablement la conséquence d'une adaptation de ces espèces, vivants au sein d'environnements pollués par des mélanges complexes, et ce pour dégrader rapidement et efficacement des mélanges de molécules proches structurellement (Stolz, 2009). En effet, les études réalisées par Romine en 1999 montrent une croissance optimale de la souche *Sphingomonas aromaticivorans* uniquement en présence d'un mélange de HAP, la vitesse de dégradation étant la plus importante uniquement dans ces conditions, où les gènes semblent tous induits (et certains probablement plus fortement) (Romine *et al.*, 1999a).

Les résultats d'expression du gène *xylE* indiquent aussi que la souche *Sphingomonas paucimobilis* sp. EPA505 utilise la voie de clivage dite *meta* et non celle dite *ortho*. En effet, ce gène est spécifique de la première voie et les gènes intervenant dans la seconde sont quant à eux absents ou non exprimés. Cependant, certaines souches bactériennes possèdent les deux voies basses *meta* et *ortho* de dégradation dont l'expression est induite différentiellement en fonction du type de substrat. Par exemple, c'est le cas chez *Rhodococcus* sp. DK17 (Kim *et al.*, 2002), et *Pseudomonas putida* F1 qui utilisent spécifiquement le clivage *ortho* pour le catabolisme du benzène (Vandecasteele, 2005). Il est donc possible que la voie *ortho* soit présente pour la souche EPA505, et spécifiquement exprimée en réponse à la présence d'autres composés que ceux testés, comme le benzène ou le biphényle, car plus adaptées à leur dégradation.

Enfin, deux des régulateurs putatifs ciblés (BphR codé par le gène *bphR*, et Orf597 codé par le gène Orf597), sont eux aussi différentiellement exprimés en fonction de la source de carbone présente. Ainsi, le gène *bphR* est fortement exprimé en présence de glucose et à la limite du seuil de détection, en présence du mélange des deux HAP. Ce gène pourrait potentiellement coder pour un répresseur des gènes codant les enzymes de dégradation de composés aromatiques lorsque ceux-ci ne sont pas disponibles pour la souche (Romine *et al.*, 1999b; Stolz, 2009). Le gène appelé Orf597, codant un régulateur putatif de la famille MucR (Romine *et al.*, 1999b), est quant à lui exprimé en présence de phénanthrène, et non en présence de fluoranthène. Il est connu que les régulateurs de la famille MucR sont impliqués dans la régulation des gènes impliqués dans la formation d'exopolysaccharides au sein des parois bactériennes, comme chez *Sinorhizobium meliloti* (Bahlawane *et al.*, 2008). La souche *Sphingomonas paucimobilis* sp. EPA505 est elle aussi capable (comme beaucoup d'espèces du genre *Sphingomonas*) à la fois de sécréter des exopolysaccharides (nommés sphingans) servant à solubiliser les HAP, mais aussi de former des structure de type biofilm (Johnsen et Karlson, 2004; Stolz, 2009). L'hypothèse émise dans ce cas est que cet activateur soit impliqué dans la réponse à la présence de phénanthrène, en agissant soit sur la production de ces sphingans, pour faciliter la solubilisation de ce HAP, ou pour assurer la formation de biofilms pour garantir la survie de la souche au sein de l'écosystème pollué et potentiellement toxique.

5. Exploration des capacités métaboliques microbiennes d'un sol pollué

L'amélioration des conditions de bioremédiation de sols pollués nécessite à la fois d'appréhender la structure des communautés microbiennes présentes, mais également de caractériser leurs potentialités métaboliques contribuant à la restauration de l'écosystème perturbé (Andreoni et Gianfreda, 2007; Stenuit *et al.*, 2008). Pour ce genre d'étude, les biopuces ADN, taxonomiques ou fonctionnelles, développées dans l'équipe, s'avère être les outils parmi les plus rapides et les plus efficaces pour caractériser ces communautés et leurs potentialités épuratrices (Joux *et al.*, 2010).

Parallèlement à l'approche biopuce taxonomique, une approche préliminaire de clonage-séquençage a été menée. Les premiers résultats sur l'étude de la diversité, présente au sein du site pollué d'intérêt, révèlent l'existence d'une grande richesse taxonomique. Ainsi, d'après les indices de diversité calculés à partir des données de la banque de clones d'ADNr

16S, le nombre total de genres identifiés reste bien en deçà de la diversité potentielle totale de ce site pollué (estimée à 345 et 325 respectivement avec les indices ACE et Chao1). Sans optimisation préalable, l'approche utilisant la biopuce taxonomique avec de l'ADN génomique comme cible, assure la détection de 45 genres, traduisant de bonnes capacités d'évaluation de la diversité. Un tel résultat montrant de meilleures résultats par l'approche biopuce ADN que par le clonage/séquençage a déjà été constaté dans d'autres études (DeSantis *et al.*, 2007; Rastogi *et al.*, 2010). Dans notre étude, le nombre de genres détectés reste faible, au vu de l'estimation du nombre total d'OTUs suite à la caractérisation de la banque de clones. Il est donc nécessaire d'optimiser les conditions d'hybridation (quantité d'ADN, température d'hybridation…). Il serait également intéressant d'hybrider les produits PCR utilisés pour la construction de la banque ce qui permettrait une réelle comparaison des deux approches. En effet, la limite de détection des biopuces pour les genres faiblement représentés au sein de l'environnement pourrait être améliorée grâce aux produits PCR (enrichissement des cibles d'intérêt, tout en réduisant la complexité de l'échantillon biologique). L'utilisation de la PCR peut cependant introduire des biais dans l'évaluation de la diversité (amorces non « universelles », amplifications préférentielles de séquences dépendant de la composition en bases et de la quantité des gènes ciblés). L'idéal serait donc d'hybrider les cibles les plus diversifiées possibles (ADNg, produits PCR, ARNr) pour obtenir une vision la plus réaliste possible (DeSantis *et al.*, 2007; Rastogi *et al.*, 2010).. L'approche restreinte de clonage-séquençage a effectivement révélé un faible nombre d'OTUs majoritaires, les autres OTUs n'étant représentés que par une séquence unique.

Les 45 genres identifiés avec la biopuce taxonomique appartiennent à cinq *phyla* : les *Actinobacteria*, les *Bacteroidetes*, les *Firmicutes*, les *Proteobacteria* et les *Spirochaetes*. Certains d'entre eux sont connus comme possédant des genres pour lesquelles des espèces impliquées dans la dégradation de polluants aromatiques (comme les genres *Mycobacterium, Gordonia, Rhodococcus, Cornyebacterium, Brevibacillus, Aeromonas, Pseudoalteromonas, Halomonas* et *Roseovarius*) ont été caractérisées (Lebkowska *et al.*, 1995; Cassidy et Hudak, 2002; Andreoni *et al.*, 2004; Seo *et al.*, 2009; Al-Mailem *et al.*, 2010; Vila *et al.*, 2010; Zanaroli *et al.*, 2010). L'approche clonage-séquençage a, quant à elle, mise en évidence la présence d'autres phyla (*Acidobacteria*, *Planctomycetes, Gemmatimonadetes*, *Chloroflexi* et *Nitrospirae*). De plus, d'autres genres (*Acidovorax, Ralstonia* et *Pseudomonas*), dont certaines espèces sont impliquées dans la dégradation de divers HAP, ont aussi été identifiés par cette approche (Seo *et al.*, 2009).

Concernant le genre dominant *Pseudomonas*, mis en évidence dans la banque de clones, il faut également préciser que celui-ci n'était pas pris en compte avec la version de Phylarray utilisée pour la détermination des sondes pour cette biopuce. Cependant, de nouvelles améliorations sur l'algorithme, assurant le traitement d'un grand nombre de séquences pour un genre donné (ce qui est le cas pour le genre *Pseudomonas*), ont permis la détermination de sondes pour ce genre, qui ont été intégrées à la nouvelle biopuce en cours d'élaboration.

Pour conclure, l'utilisation complémentaire de ces deux techniques reste donc essentielle, afin de relier les données obtenues par les deux approches (Rastogi *et al.*, 2010). Cependant, les moyens humains et financiers nécessaires à l'obtention d'un grand nombre de séquences par des techniques classiques pour une analyse plus fine, nous fait nous tourner vers l'utilisation de techniques de séquençage dites haut-débit (comme le pyroséquençage) (Roesch *et al.*, 2007). Ces nouvelles méthodes permettent d'obtenir plus de 25 000 séquences rapidement, et sont de plus en plus utilisées pour étudier la diversité d'écosystèmes complexes (Campbell *et al.*, 2009; Youssef *et al.*, 2009). Leur principal défaut reste cependant l'obtention de séquences assez courtes (entre 400 et 500 bases actuellement) nécessitant une reconstruction des contigs fiable et efficace.

L'utilisation de la biopuce ADN fonctionnelle exploratoire a permis de caractériser les potentialités métaboliques de dégradation de polluants aromatiques au sein du site étudié, donnant un aperçu plus global que des approches classiques (Leys *et al.*, 2004; Kim et Crowley, 2007a). Trente et un gènes ciblés avec cette biopuce ont pu être mis en évidence. Les voies de dégradation du phénanthrène et du naphtalène ont ainsi été détectées (grâce aux signaux des sondes ciblant les gènes *phnA2a, ahdA1c, ahdA2c, bphC, nahD, bphB, bphA3, ahdA4 et phdI* pour les voies du phénanthrène, et *nahF* pour la voie du naphtalène) (Kim *et al.*, 1997; Iwabuchi et Harayama, 1998b; Romine *et al.*, 1999b; Pinyakong *et al.*, 2003b; Demaneche *et al.*, 2004; Cho *et al.*, 2005; Keck *et al.*, 2006; Ní Chadhain *et al.*, 2007; Stolz, 2009) (Figure 54). Il a également été possible de détecter des gènes impliqués dans la voie de dégradation du gentisate (*nagG, nagH* et *nagR*), et un gène de la voie de clivage dite *ortho* (*catA*) (Zhou *et al.*, 2001; Vandecasteele, 2005). L'absence de détection du gène *xylE* laisse supposer l'absence de la voie de clivage *meta*. Cependant, cette voie pourrait exister, car il a été démontré que BphC (codé par le gène *bphC*) pouvait également cliver le catéchol, et donc remplacer l'enzyme XylE (Kim et Zylstra, 1999). Enfin, les voies de dégradation de certains composés aromatiques ont été partiellement identifiées (à travers la présence des gènes *xylA,*

xylM, xylY et *bphF*), bien que d'autres gènes intervenant dans ces voies n'aient pas été détectés (comme *xylX* et *xylC*). L'ensemble de ces résultats démontre que les communautés microbiennes de l'écosystème étudié possèdent les capacités métaboliques de dégradation des HAP et/ou d'autres molécules aromatiques.

Enfin, plusieurs régulateurs putatifs, ciblés pour leur implication potentielle dans la régulation des processus biologiques de dégradation des composés aromatiques, ont été mis en évidence, comme l'Orf597, qui code pour un régulateur putatif de la famille MucR. Cette famille de régulateurs est impliquée dans la régulation de gènes codant des protéines permettant la formation de polysaccharides au sein des parois bactériennes, décrit chez *Sinorhizobium meliloti*. Ces polysaccharides servent par exemple à la fixation des cellules bactériennes sur la racine de la plante (Bahlawane *et al.*, 2008). Or, il est connu que certains genres (comme *Sphingomonas* ou *Pseudomonas*) produisent ce type de composés (excrétés ou non) pour mieux solubiliser les HAP, comme nous l'avons déjà évoqué précédemment (Andreoni *et al.*, 2004; Johnsen et Karlson, 2004; Stolz, 2009).

L'impossibilité de pouvoir détecter des gènes codant pour la sous-unité α de la dioxygénase initiale par l'approche biopuce, alors que ces derniers sont présents, comme l'a montré l'amplification PCR, révèle que les sondes déterminées ne semblent pas suffisamment exploratoires. En effet, les séquences environnementales isolées par PCR, bien que codant pour des protéines correspondant à la même famille enzymatique assurant l'attaque initial du HAP, sont trop divergentes des sondes sensées les mettre en évidence. Il est donc probable que les espèces possédant ce gène sont phylogénétiquement éloignées de celles (appartenant principalement aux genres *Sphingomonas* et *Pseudomonas*) dont les séquences ont été sélectionnées pour l'élaboration des sondes (Baek *et al.*, 2009). Ce résultat confirme donc que les sondes déterminées, bien qu'étant spécifiques du gène *phnA1a* (codant pour la sous-unité α de la dioxygénase initiale), restent trop restrictives pour assurer l'identification d'autres gènes codant pour cette enzyme. Il est donc indispensable de déterminer de nouvelles sondes ciblant toutes les classes oxydatives de sous-unités α de dioxygénases initiales connues (Kweon *et al.*, 2008).

L'aspect exploratoire reste malgré tout un avantage majeur des sondes définies avec le logiciel Metabolic Design. Ainsi, pour plusieurs autres gènes ciblés, plusieurs variants potentiels ont pu être répertoriés. Pour confirmer l'importance de cet aspect exploratoire sur l'évaluation de la diversité génique de l'environnement étudié, une recherche de similarité de séquences pour les sondes positives ciblant le gène *phnA2a* a été réalisée en utilisant

l'algorithme BLASTn, et la banque de données EMBL. Cette étude a révélé que les séquences de 21 des 37 sondes positives montrent de fortes identités (avec seulement 0, 1 ou 2 mésappariements) avec d'autres séquences du gène *phnA2a* (GU441603, EF152282, EU024112, EU526899 et AJ633551), isolées chez des espèces comme *Novosphingobium* sp. H25, *Cycloclasticus* sp. NY93E ou *Sphingomonas* sp. CHY-1. Les autres sondes donnant un signal positif ne montrent quant à elles que de faibles identités (ayant plus de 2 mésappariements) avec les séquences du gène *phnA2a* disponibles. Les sondes ont donc vraisemblablement permis la détection de nouveau variants, dont les séquences complètes pourraient être déterminées par une approche d'amplification PCR et de clonage-séquençage, ou de piégeage de gènes sur biopuce.

De même, il est important de noter la forte diversité potentiellement détectée (168 sondes sur 384 ciblant une même région de ce gène donnent un signal supérieur au seuil défini) pour le gène *phdI*, codant une enzyme spécifique d'une voie annexe (appelée voie de Kiyohara) décrite chez les genres *Nocardioides* et *Mycobacterium* (Iwabuchi et Harayama, 1998b; Kim *et al.*, 2007; Kweon *et al.*, 2007). Ces résultats confirment de nouveau l'aspect exploratoire de notre approche, qui permet d'avoir un aperçu plus exhaustif sur la diversité génique des gènes ciblées, contrairement aux autres approches de détermination de sondes jusqu'à présent développées, où seuls les gènes connus sont ciblés (Rhee *et al.*, 2004; Liang *et al.*, 2009a; Liang *et al.*, 2009b).

Ces deux approches de biopuces taxonomique et fonctionnelle, menées conjointement, nous permettent d'établir plusieurs liens entre la structure des communautés bactériennes, et les processus biologiques potentiellement mis en jeu dans la dégradation des HAP.

Les sous-unités α de dioxygénases initiales sont une famille de protéines assez large, dont les séquences peuvent être assez divergentes d'un point de vue séquence, selon leur spectre de substrat, et l'organisme chez qui elles ont été isolées, et sont donc très étudiées (Vandecasteele, 2005; Kim et Crowley, 2007b; Bordenave *et al.*, 2008; Cébron *et al.*, 2008; Kweon *et al.*, 2008). Les sondes développées pour *phnA1a* ciblent plusieurs gènes codant pour ces protéines, mais principalement affiliés aux genres *Sphingomonas* et *Pseudomonas*. L'impossibilité à détecter ce gène *phnA1a* avec les sondes développées montre donc l'absence de ce dernier au sein de l'écosystème étudié (Kweon *et al.*, 2008; Baek *et al.*, 2009), bien que les espèces du genre *Pseudomonas* soient fortement représentés au sein de l'écosystème. Cependant, l'approche PCR mise en place a montré l'existence de séquences codant pour des sous-unités α de dioxygénases (proches de *dbtAc* et de *nagAc* isolés notamment chez des

espèces des genres *Burkholderia* et *Polaromonas*) au sein de cet écosystème. La présence d'espèces appartenant aux genres *Mycobacterium, Rhodococcus* et *Pseudomonas*, identifiés lors de l'étude la structure des communautés microbiennes du site pollué, laisse également supposer la présence d'autres sous-unités α de dioxygénases, plus proches de *nidA*, *nidA3*, *padAa2* ou *phtAa*, mais non ciblées sur la biopuce. En effet, chez ces espèces, ces gènes codent des protéines ayant la même fonction biologique que PhnA1a (Larkin *et al.*, 1999; Khan *et al.*, 2001; Krivobok *et al.*, 2003; Kim *et al.*, 2007).

Parmi les espèces appartenant au genre *Pseudomonas,* certaines sont néanmoins connues pour être impliquées dans la dégradation de plusieurs HAP (Vandecasteele, 2005; Stolz, 2009). Les nombreux gènes de sous-unités d'oxygénases détectés par les sondes de la biopuce fonctionnelle pourraient donc appartenir au patrimoine génétique de ces espèces. De plus, la voie de clivage dite *ortho* (identifiée par la présence du gène *catA*) a aussi été décrite chez ce genre (Rothmel *et al.*, 1990). L'identification de ce gène peut laisser penser qu'il est issu de certaines de ces espèces affiliées à ce genre. Comme le suggère l'absence de l'enzyme impliquée dans l'attaque initiale, il est bien sûr envisageable que ces bactéries du genre *Pseudomonas* utilisent certains métabolites provenant de la dégradation réalisée par des espèces appartenant aux genres *Mycobacterium* ou *Rhodococcus*. Finalement, comme certaines espèces de ce genre sont capables de solubiliser les polluants géosorbés au sein de sols pollués (Andreoni *et al.*, 2004), l'Orf597 identifié pourrait provenir de ces bactéries. En effet, celui-ci code un régulateur putatif de la famille MucR, qui agit sur la production de polysaccharides (excrétés ou liés à la paroi bactérienne), ou de biosurfactants, pour faciliter la solubilisation des HAP (Cheng *et al.*, 2004; Bahlawane *et al.*, 2008). Ce gène pourrait également avoir pour origine des espèces du genre *Rhodococcus*, aussi capable de s'adsorber aux composés aromatiques non solubilisés (Andreoni *et al.*, 2004).

Au sein d'écosystèmes complexes, les microorganismes fonctionnent généralement en *consortia*. C'est également le cas des microorganismes impliqués dans les processus de biodégradation des HAP (Molina *et al.*, 2009; Vila *et al.*, 2010; Zanaroli *et al.*, 2010). Ainsi, les genres *Gordonia* (Zanaroli *et al.*, 2010), *Cornyebacterium* (Cassidy et Hudak, 2002)*, Brevibacillus, Pseudoalteromonas* et *Halomonas* (Andreoni *et al.*, 2004; Al-Mailem *et al.*, 2010), retrouvés au sein du site pollué, ont déjà été caractérisés au sein de *consortia* impliqués dans la dégradation d'hydrocarbures (aromatiques ou non). Ces mêmes *consortia* regroupent également les genres *Mycobacterium, Pseudomonas* et/ou *Rhodococcus*, eux aussi mis en évidence au sein du site pollué. Le fonctionnement de ce *consortia* pourrait se baser sur la capacité de solubilisation des HAP les plus récalcitrants par des espèces appartenant aux

genres *Pseudomonas* et/ou *Rhodococcus* (Andreoni *et al.*, 2004; Cheng *et al.*, 2004). Cette solubilisation permettrait aux espèces capables de dégrader ces HAP récalcitrants (comme celles du genre *Mycobacterium*), d'accéder facilement à ces composés et donc d'agir plus efficacement. Ces dernières espèces fourniraient en retour certains métabolites provenant de la dégradation initiale des HAP aux espèces ayant solubilisé les composés. Cette hypothèse de fonctionnement reste bien sûr à vérifier par d'autres analyses, comme par exemple par un enrichissement des communautés épuratrices de ce sol contaminé et une caractérisation fine de leur fonctionnement et de leurs capacités. Une approche intéressante à mettre en place pour étudier ces communautés microbiennes épuratrices serait celle de la technique appelée Stable Isotope Probing (ou SIP) (Sul *et al.*, 2009). Elle consiste en l'apport d'un substrat d'intérêt (dans notre cas le phénanthrène) marqué au C_{13}. Celui-ci est assimilé par certains microorganismes, et incorporé au niveau des acides nucléiques grâce aux réactions d'anabolisme. Le neutron supplémentaire de l'isotope stable rend les constituants cellulaires plus lourds que ceux ayant incorporé l'atome naturel à la place. Ainsi, une fois extraits, les acides nucléiques des microorganismes ayant utilisé le substrat marqué peuvent être séparés et analysés de ceux des microorganismes ne l'ayant pas consommé.

Il reste bien sûr envisageable que les communautés fongiques aient également un rôle dans la dégradation des HAP, car certaines sont connues pour posséder des capacités épuratrices, notamment vis-à-vis de ce type de composés (Cerniglia, 1997; Haritash et Kaushik, 2009; Zanaroli *et al.*, 2010). De même, les communautés anaérobies, de plus en plus étudiées au sein de sols pollués ont aussi un rôle prépondérant dans ces dégradations (Chang *et al.*, 2006; Chang *et al.*, 2008; Fuchedzhieva *et al.*, 2008; Kim *et al.*, 2008; Haritash et Kaushik, 2009; Li *et al.*, 2009a). Pour prendre en compte ces populations dans la compréhension globale du fonctionnement de l'écosystème, il est donc indispensable de développer des biopuces permettant d'apprécier la présence de ces populations de microorganismes, et d'évaluer leurs potentialités métaboliques épuratrices. Des techniques complémentaires peuvent également être utilisées, comme le SIP, qui peut par exemple être couplé à des techniques de séquençage comme la métagénomique (Suenaga *et al.*, 2009; Chen et Murrell, 2010; Chen *et al.*, 2010). Les données obtenues peuvent ainsi permettre de relier la structure des communautés aux fonctions métaboliques mises en jeu dans la dégradation des HAP. La capacité à lier la structure à la fonction demeure donc un des objectifs majeurs de l'écologie microbienne.

CONCLUSION GENERALE ET PERSPECTIVES

CONCLUSION GENERALE ET PERSPECTIVES

Les objectifs de cette thèse ont donc été de développer, de valider et d'utiliser une biopuce exploratoire, capable d'appréhender, au sein d'un écosystème pollué, la diversité génique codant les protéines impliquées dans les processus de dégradation de composés aromatiques, et plus particulièrement de HAP. En effet, les biopuces ADN fonctionnelles utilisées actuellement ne permettent que de cibler les gènes dont les séquences ont été caractérisées. Cette limitation est liée au fait qu'il n'existe pas encore de logiciels pour définir des sondes exploratoires, assurant ainsi la prise en compte de toute la diversité génique. En effet, ce type de sondes permettrait d'obtenir un aperçu complet des processus biologiques mis en jeu au sein des écosystèmes d'intérêt.

C'est dans ce but que l'on a développé un outil informatique, nommé Metabolic Design, dédié : (i) à la reconstruction de processus biologiques, *in silico*, via la fouille de données de banques publiques et/ou personnelles ; (ii) à la détermination de sondes exploratoires pour des biopuces fonctionnelles, et (iii) à l'évaluation de leur spécificité. L'utilisation de Metabolic Design a ainsi permis de définir 72 sondes dégénérées, ciblant 40 gènes différents, dont les produits sont majoritairement impliqués dans les voies métaboliques de dégradation de molécules aromatiques.

Cette biopuce fonctionnelle a ensuite été validée en se basant sur l'étude de la souche *Sphingomonas paucimobilis* sp. EPA505, connue pour dégrader différents HAP. Cette étude a tout d'abord confirmé l'aspect exploratoire et spécifique des sondes développées sans *a priori*, en comparant les séquences des sondes spécifiques donnant les plus forts signaux, aux séquences caractérisées de la souche modèle. De plus, le suivi de l'expression des gènes de la souche modèle, par des approches de PCR quantitative et d'hybridation de biopuces ADN, a montré que les sondes développées sont suffisamment sensibles pour réaliser un suivi semi-quantitatif de l'expression des gènes étudiés.

D'autre part, les résultats d'expression obtenus pour la souche modèle apportent de nouvelles données pour une meilleure compréhension des voies métaboliques impliquées dans la dégradation des HAP, et de leurs régulations. En effet, en présence d'un mélange de HAP, de nombreux gènes sont fortement exprimés. Or, pour le genre *Sphingomonas*, les gènes codant des enzymes impliquées dans la dégradation des HAP (souvent impliquées dans plusieurs voies différentes), sont organisés au sein de mêmes clusters. Cette surexpression serait potentiellement la conséquence d'une adaptation de ces espèces, vivants au sein

d'environnements pollués par des mélanges complexes, pour dégrader rapidement et efficacement des mélanges de molécules proches structurellement.

L'aspect exploratoire des sondes développées a également été confirmé par les résultats obtenus lors de l'étude d'un écosystème de type sol, contaminé par des HAP. En effet, cette analyse a permis d'évaluer la diversité génique des gènes ciblés, et donc les potentialités métaboliques épuratrices présentes, au sein de cet écosystème d'intérêt. De plus, ces résultats ont été corrélés avec les données préliminaires phylogénétiques rassemblées par des approches d'amplification PCR et d'hybridation d'une biopuce taxonomique. Ainsi, avec la biopuce fonctionnelle développée, le gène *phdI* a été détecté. Ce dernier code pour une 1-hydroxy-2-naphthoate dioxygénase, spécifiquement décrite chez les genres *Arthrobacter, Nocardioides* et *Mycobacterium*, ce dernier genre ayant été détecté en parallèle avec les analyses taxonomiques. Or, les résultats pour ce gène ont montré une grande diversité génique avec la biopuce exploratoire, non encore décrite au sein des bases de données, où seules quelques séquences ont été caractérisées. Ces résultats démontrent donc l'intérêt de notre approche pour évaluer les potentialités métaboliques de l'écosystème étudié, dans leur totalité, en explorant également la fraction inconnue des microorganismes présents.

De plus, la complémentarité des deux biopuces ADN : fonctionnelle et taxonomique, est cruciale pour l'analyse d'écosystèmes dans leur globalité. En effet, la corrélation précise des données de structures et de fonctions obtenues permettrait de réellement appréhender les mécanismes contribuant au fonctionnement et aux adaptations des communautés microbiennes au sein d'un environnement perturbé. D'autres techniques peuvent permettre de relier la structure des communautés aux fonctions biologiques, comme le SIP, couplé à des techniques de séquençage (par exemple la métagénomique) (Suenaga *et al.*, 2009; Chen et Murrell, 2010; Chen *et al.*, 2010).

Ces travaux apportent également de nouvelles perspectives de recherche, tant sur l'étude de la souche modèle, que sur l'approche mise en place. D'ores et déjà, pour la souche EPA505, les premiers éléments obtenus nous permettent de proposer des actions pour mieux appréhender les mécanismes, et les acteurs mis en jeu pour assurer la régulation de l'expression des gènes codant des enzymes impliquées dans les différentes voies de dégradation des HAP. Ainsi, une étude, par des approches de gel retard ou de DNase I footprinting, de la zone non codante comprise entre les gènes *xylX* et *bphC*, et plus particulièrement de la région conservée dont la structure semble palindromique, permettrait d'évaluer son implication dans la régulation de ces gènes. En parallèle, la caractérisation de la séquence précise des gènes *bphR* et *orf597*, et de leurs produits pour la souche *Sphingomonas*

paucimobilis sp. EPA505 permettrait d'évaluer l'implication potentielle de ces régulateurs dans les voies de dégradation de ce type de molécules. Il serait également intéressant de réaliser une analyse cristallographique de l'enzyme PhnA1a afin de vérifier si les acides aminés différents, par rapport à PhnI issue de la souche CHY-1, sont impliqués dans la reconnaissance au substrat. Une autre approche envisageable pour étudier le mode de fonctionnement de cette enzyme serait de réaliser de la mutagenèse sur ces résidus, puis d'en évaluer les impacts dans sa reconnaissance des substrats et dans son activité enzymatique.

Bien que l'utilisation de Metabolic Design ait permis de mettre en évidence les voies métaboliques de dégradation de HAP et leurs régulations chez la souche modèle, l'amélioration de cet outil informatique est essentielle. En effet, prendre en compte de nouveaux paramètres permettrait d'améliorer le design des sondes. Il serait ainsi nécessaire de considérer des critères thermodynamiques et les calculs de Tm pour la sélection des sondes (Pozhitkov *et al.*, 2006; Pozhitkov *et al.*, 2007; Li *et al.*, 2008; Arslan et Laurenzi, 2009; Mueckstein *et al.*, 2010). L'étude d'un écosystème complexe a également permis de montrer la présence de nombreux gènes, ou de groupes de gènes, dont les produits sont potentiellement impliqués dans la dégradation de HAP. L'isolement, par exemple par amplification PCR, des gènes détectés par les sondes développées, permettrait de déterminer la réelle diversité génique dans cet écosystème. En parallèle, il serait important de définir des sondes pour évaluer toute la diversité génique des gènes codant la sous-unité α de la dioxygénase initiale, biomarqueur fonctionnel important des microorganismes épurateurs. Pour cela, il est indispensable de se baser sur les données de classification disponibles, comme la banque OxDBase, spécifique des oxygénases impliquées dans les processus de biodégradation (Kweon *et al.*, 2008; Arora *et al.*, 2009). Enfin, bien que certains gènes aient été détectés via l'utilisation de la biopuce ADN exploratoire, leur expression n'a pu être confirmée. Il serait donc crucial de mettre en place une méthode pour extraire les ARNs de cet écosystème pollué, afin de véritablement évaluer les voies métaboliques actives de ce sol pollué.

De par l'existence de très nombreux polluants (comme les métaux lourds, les solvants organiques ou les pesticides) retrouvés au sein des écosystèmes, l'outil Metabolic Design pourrait être utilisé pour définir des sondes afin d'étudier les processus biologiques mis en jeu par les microorganismes (champignons, algues, bactéries et archées) pour assurer leurs épurations. Il serait donc intéressant de cibler les voies métaboliques de dégradation de pesticides ou de polluants organiques, mais également les voies de réduction des métaux lourds, comme l'arsenic par exemple (Achour-Rokbani *et al.*, 2007; Achour-Rokbani *et al.*,

2010). Metabolic Design permettant de définir des sondes pour n'importe quel type de gène, il serait également judicieux d'étudier, en plus de la métabolisation des polluants, leur accessibilité aux microorganismes, par exemple par l'analyse des gènes codant les protéines de transports de ces molécules. Il serait également très intéressant de coupler cette biopuce à des techniques comme le SIP, apportant à la fois des informations sur les communautés microbiennes actives (si l'on utilise l'ARN total extrait de l'écosystème étudié, marqué classiquement), et des données sur celles réellement impliquées dans la dégradation des composés étudiés (grâce à l'incorporation dans leur ARN de bases marquées au C_{13}, venant de la minéralisation des composés, détectées par microradiographie) (Neufeld *et al.*, 2007).

Tout en assurant la mise en évidence des potentialités métaboliques impliquées au cours des processus de bioremédiation, le suivi du niveau d'expression des gènes avec les biopuces ADN, permettrait de comprendre les régulations régissant le fonctionnement de ces processus. L'ensemble de ces connaissances faciliteraient par la suite les optimisations des potentialités épuratrices des microorganismes (Liang *et al.*, 2009b). Il serait ainsi possible de lever certaines régulations, avec l'ajout d'effecteurs identifiés, ou alors d'introduire des microorganismes exogènes possédant des capacités métaboliques absentes, et indispensables pour dépolluer l'écosystème d'intérêt.

L'utilisation du logiciel Metabolic Design peut également s'étendre à des domaines de recherches autre que la bioremédiation (Dharmadi et Gonzalez, 2004; Bansal, 2005; Singh, 2010). En effet, les microorganismes sont également une source de produits à haute valeur ajoutée, comme les biopesticides, les antibiotiques, les molécules de chimie fine, les agents anti-parasites, les acides aminés, les vitamines, les détergents, etc. qui représentent un marché de plusieurs dizaines de milliards d'euros (Singh, 2010). L'étude des processus biologiques mis en jeu dans la production de ces molécules d'intérêt pourrait permettre d'optimiser les procédés de production. De plus, la recherche de ces voies métaboliques au sein d'environnements complexes pourrait permettre l'identification de nouveaux microorganismes, ou de nouveaux variants, associés à ces métabolismes de haute valeur ajoutée.

Ainsi, nos résultats et les approches que nous envisageons ouvrent des perspectives intéressantes, quant à l'étude du fonctionnement global des communautés microbiennes au sein de différents écosystèmes, ces communautés étant une source presque inépuisable de capacités métaboliques encore largement méconnues.

REFERENCES

REFERENCES

Acar, Y. B. and A. N. Alshawabkeh (1993). "Principles of electrokinetic remediation." *Environmental Science & Technology* **27**(13): 2638-2647.

Achour-Rokbani, A., P. Bauda and P. Billard (2007). "Diversity of arsenite transporter genes from arsenic-resistant soil bacteria." *Research in Microbiology* **158**(2): 128-137.

Achour-Rokbani, A., A. Cordi, P. Poupin, P. Bauda and P. Billard (2010). "Characterization of the *ars* Gene Cluster from Extremely Arsenic-Resistant *Microbacterium* sp. Strain A33." *Applied and Environmental Microbiology* **76**(3): 948-955.

Adler, P., J. Reimand, J. Janes, R. Kolde, H. Peterson and J. Vilo (2008). "KEGGanim: pathway animations for high-throughput data." *Bioinformatics* **24**(4): 588-590.

Adriaens, M. E., M. Jaillard, A. Waagmeester, S. L. M. Coort, A. R. Pico and C. T. A. Evelo (2008). "The public road to high-quality curated biological pathways." *Drug Discovery Today* **13**(19-20): 856-862.

Ahn, C. K., Y. M. Kim, S. H. Woo and J. M. Park (2008). "Soil washing using various nonionic surfactants and their recovery by selective adsorption with activated carbon." *Journal of Hazardous Materials* **154**(1-3): 153-160.

Ahn, Y., H. Jung, R. Tatavarty, H. Choi, J.-w. Yang and I. S. Kim (2005). "Monitoring of petroleum hydrocarbon degradative potential of indigenous microorganisms in ozonated soil." *Biodegradation* **16**(1): 45-56.

Al-Mailem, D. M., N. A. Sorkhoh, M. Marafie, H. Al-Awadhi, M. Eliyas and S. S. Radwan (2010). "Oil phytoremediation potential of hypersaline coasts of the Arabian Gulf using rhizosphere technology." *Bioresource Technology* **101**(15): 5786-5792.

Altermann, E. and T. Klaenhammer (2005). "PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database." *BMC Genomics* **6**(1): 60.

Altschul, S., W. Gish, W. Miller, E. Myers and D. Lipman (1990). "Basic local alignment search tool." *Journal of Molecular Biology* **215**(3): 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Research* **25**(17): 3389-3402.

Amann, R. and W. Ludwig (2000). "Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology." *FEMS Microbiology Reviews* **24**(5): 555-565.

Andreoni, V., L. Cavalca, M. A. Rao, G. Nocerino, S. Bernasconi, E. Dell'Amico, M. Colombo and L. Gianfreda (2004). "Bacterial communities and enzyme activities of PAHs polluted soils." *Chemosphere* **57**(5): 401-412.

Andreoni, V. and L. Gianfreda (2007). "Bioremediation and monitoring of aromatic-polluted habitats." *Applied Microbiology and Biotechnology* **76**(2): 287-308.

Arcuri, H., G. Zafalon, E. Marucci, C. Bonalumi, N. da Silveira, J. Machado, W. de Azevedo and M. Palma (2010). "SKPDB: a structural database of shikimate pathway enzymes." *BMC Genomics* **11**(1): 12.

Arora, P., M. Kumar, A. Chauhan, G. Raghava and R. Jain (2009). "OxDBase: a database of oxygenases involved in biodegradation." *BMC Research Notes* **2**(1): 67.

Arslan, E. and I. Laurenzi (2009). "An efficient algorithm for the stochastic simulation of the hybridization of DNA to microarrays." *BMC Bioinformatics* **10**(1): 411.

Ashburner, M. and S. Lewis (2002). "On ontologies for biologists: the Gene Ontology--untangling the web." *Novartis Found Symposium* **247**: 66-80.

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones and A. J. Weightman (2006). "New Screening Software Shows that Most Recent Large 16S rRNA Gene Clone

Libraries Contain Chimeras." *Applied and Environmental Microbiology* **72**(9): 5734-5741.

Attwood, T. K. (2002). "The PRINTS database: A resource for identification of protein families." *Briefings in Bioinformatics* **3**(3): 252-263.

Bader, G. D., M. P. Cary and C. Sander (2006). "Pathguide: a Pathway Resource List." *Nucleic Acids Research* **34**(suppl_1): D504-506.

Baek, S., O. Kweon, S.-J. Kim, D.-H. Baek, J. J. Chena and C. E. Cerniglia (2009). "ClassRHO: A platform for classification of bacterial rieske non-heme iron ring-hydroxylating oxygenases." *Journal of Microbiological Methods* **76**(3): 307-309.

Bahlawane, C., M. McIntosh, E. Krol and A. Becker (2008). "*Sinorhizobium meliloti* Regulator MucR Couples Exopolysaccharide Synthesis and Motility." *Molecular Plant-Microbe Interactions* **21**(11): 1498-1509.

Baker, G. C., J. J. Smith and D. A. Cowan (2003). "Review and re-analysis of domain-specific 16S primers." *Journal of Microbiological Methods* **55**(3): 541-555.

Bansal, A. (2005). "Bioinformatics in microbial biotechnology - a mini review." *Microbial Cell Factories* **4**(1): 19.

Barbe, V., D. Vallenet, N. Fonknechten, A. Kreimeyer, S. Oztas, L. Labarre, S. Cruveiller, C. Robert, S. Duprat, P. Wincker, L. N. Ornston, J. Weissenbach, P. Marliere, G. N. Cohen and C. Medigue (2004). "Unique features revealed by the genome sequence of *Acinetobacter* sp. ADP1, a versatile and naturally transformation competent bacterium." *Nucleic Acids Research* **32**(19): 5766-5779.

Barnsley, E. A. (1976). "Role and regulation of the *ortho* and *meta* pathways of catechol metabolism in pseudomonads metabolizing naphthalene and salicylate." *Journal of Bacteriology* **125**(2): 404-408.

Basta, T., A. Keck, J. Klein and A. Stolz (2004). "Detection and Characterization of Conjugative Degradative Plasmids in Xenobiotic-Degrading Sphingomonas Strains." *The Journal of Bacteriology* **186**(12): 3862-3872.

Batie, C. J., D. P. Ballou and C. C. Correll (1992). **Phthalate dioxygenase reductase and related flavin-iron-sulfur containing electron transferases.**, CRC Press.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2009). "GenBank." *Nucleic Acids Research* **37**(suppl_1): D26-31.

Biache, C., L. Mansuy-Huault, P. Faure, C. Munier-Lamy and C. Leyval (2008). "Effects of thermal desorption on the composition of two coking plant soils: Impact on solvent extractable organic compounds and metal bioavailability." *Environmental Pollution* **156**(3): 671-677.

Boguski, M., T. Lowe and C. Tolstoshev (1993). "dbEST--database for "expressed sequence tags"." *Nature Genetics* **4**(4): 332-333.

Bonnard, M., S. Devin, C. Leyval, J. L. Morel and P. Vasseur (2010). "The influence of thermal desorption on genotoxicity of multipolluted soil." *Ecotoxicology and Environmental Safety* **73**(5): 955-960.

Bordenave, S., M. Goñi-urriza, C. Vilette, S. Blanchard, P. Caumette and R. Duran (2008). "Diversity of ring-hydroxylating dioxygenases in pristine and oil contaminated microbial mats at genomic and transcriptomic levels." *Environmental Microbiology* **10**(12): 3201-3211.

Bot, A. and J. Benites (2005). The importance of soil organic matter. FAO Soils Bulletin 80, Food and Agriculture Organization of the United Nations.

Bozdech, Z., J. Zhu, M. Joachimiak, F. Cohen, B. Pulliam and J. DeRisi (2003). "Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray." *Genome Biology* **4**(2): R9.

Brilli, M., R. Fani and P. Lio (2008). "Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes." *Briefings in Bioinformatics* **9**(1): 34-45.

Brodie, E. L., T. Z. DeSantis, D. C. Joyner, S. M. Baek, J. T. Larsen, G. L. Andersen, T. C. Hazen, P. M. Richardson, D. J. Herman, T. K. Tokunaga, J. M. Wan and M. K. Firestone (2006). "Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation." *Applied and Environmental Microbiology* **72**(9): 6288-6298.

Brouwer, R. W. W., O. P. Kuipers and S. A. F. T. van Hijum (2008). "The relative value of operon predictions." *Briefings in Bioinformatics* **9**(5): 367-375.

Bru, C., E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar and D. Kahn (2005). "The ProDom database of protein domain families: more emphasis on 3D." *Nucleic Acids Research* **33**(suppl_1): D212-215.

Brzostowicz, P. C., A. B. Reams, T. J. Clark and E. L. Neidle (2003). "Transcriptional Cross-Regulation of the Catechol and Protocatechuate Branches of the β-Ketoadipate Pathway Contributes to Carbon Source-Dependent Expression of the *Acinetobacter* sp. Strain ADP1 *pobA* Gene." *Applied and Environmental Microbiology* **69**(3): 1598-1606.

Bugg, T., J. M. Foght, M. A. Pickard and M. R. Gray (2000). "Uptake and Active Efflux of Polycyclic Aromatic Hydrocarbons by *Pseudomonas fluorescens* LP6a." *Applied and Environmental Microbiology* **66**(12): 5387-5392.

Campbell, B. J., S. W. Polson, T. E. Hanson, M. C. Mack and E. A. G. Schuur (2009). "The effect of nutrient deposition on bacterial communities in Arctic tundra soil." *Environmental Microbiology* **12**(7): 1842-1854.

Capponi, C., J. Chabalier, Y. Quentin and G. Fichant (2001). "A Knowledge Base for Integrated Biological Systems." *IEEE Intelligent Systems* **16**: 52-60.

Carbajosa, G., A. Trigo, A. Valencia and I. Cases (2009). "Bionemo: molecular information on biodegradation metabolism." *Nucleic Acids Research* **37**(suppl_1): D598-602.

Carbon, S., A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis, G. O. H. the Ami and G. the Web Presence Working (2009). "AmiGO: online access to ontology and annotation data." *Bioinformatics* **25**(2): 288-289.

Caspi, R., T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang and P. D. Karp (2010). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases." *Nucleic Acids Research* **38**(suppl_1): D473-479.

Caspi, R., H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang and P. D. Karp (2008). "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." *Nucleic Acids Research* **36**(suppl_1): D623-631.

Cassidy, D. P. and A. J. Hudak (2002). "Microorganism Selection and Performance in Bioslurry Reactors Treating PAH-Contaminated Soil." *Environmental Technology* **23**: 1033-1042.

Cébron, A., M.-P. Norini, T. Beguiristain and C. Leyval (2008). "Real-Time PCR quantification of PAH-ring hydroxylating dioxygenase (PAH-RHDα) genes from Gram positive and Gram negative bacteria in soil and sediment samples." *Journal of Microbiological Methods* **73**(2): 148-159.

Cerami, E., G. Bader, B. Gross and C. Sander (2006). "cPath: open source software for collecting, storing, and querying biological pathways." *BMC Bioinformatics* **7**(1): 497.

Cerniglia, C. E. (1992). "Biodegradation of polycyclic aromatic hydrocarbons." *Biodegradation* **3**(2-3): 351-368.

Cerniglia, C. E. (1997). "Fungal metabolism of polycyclic aromatic hydrocarbons: past, present and future applications in bioremediation." *Journal of Industrial Microbiology and Biotechnology* **19**(5): 324-333.

Cerniglia, C. E. and M. A. Heitkamp (1989). **Microbial Degradation of Polycyclic Aromatic Hydrocarbons (PAH) in the Aquatic Environment**. Baca Raton Florida, CRC Press, Inc.,.

Chang, B.-V., I. Chang and S. Yuan (2008). "Anaerobic Degradation of Phenanthrene and Pyrene in Mangrove Sediment." *Bulletin of Environmental Contamination and Toxicology* **80**(2): 145-149.

Chang, W., Y. Um and T. Holoman (2006). "Polycyclic Aromatic Hydrocarbon (PAH) Degradation Coupled to Methanogenesis." *Biotechnology Letters* **28**(6): 425-430.

Chen, F., A. J. Mackey, C. J. Stoeckert, Jr. and D. S. Roos (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." *Nucleic Acids Res.* **34**(suppl_1): D363-368.

Chen, H. and B. Sharp (2002). "Oliz, a suite of Perl scripts that assist in the design of microarrays using 50mer oligonucleotides from the 3' untranslated region." *BMC Bioinformatics* **3**(1): 27.

Chen, Y. and J. C. Murrell (2010). "When metagenomics meets stable-isotope probing: progress and perspectives." *Trends in Microbiology* **18**(4): 157-163.

Chen, Y., J. Vohra and J. Murrell (2010). "Applications of DNA-stable isotope probing in bioremediation studies." *Methods in Molecular Biology* **599**: 129-139.

Cheng, K., Z. Zhao and J. Wong (2004). "Solubilization and desorption of PAHs in soil-aqueous system by biosurfactants produced from *Pseudomonas aeruginosa* P-CG3 under thermophilic condition." *Environmental Technology* **25**(10): 1159-1165.

Cho, J.-C. and J. M. Tiedje (2001). "Bacterial Species Determination from DNA-DNA Hybridization by Using Genome Fragments and DNA Microarrays." *Applied and Environmental Microbiology* **67**(8): 3677-3682.

Cho, O., K. Y. Choi, G. J. Zylstra, Y. S. Kim, S. K. Kim, J. H. Lee, H. Y. Sohn, G. S. Kwon, Y. M. Kim and E. Kim (2005). "Catabolic role of a three-component salicylate oxygenase from *Sphingomonas yanoikuyae* B1 in polycyclic aromatic hydrocarbon degradation." *Biochemical and Biophysical Research Communications* **327**(3): 656-662.

Choi, C., R. Munch, S. Leupold, J. Klein, I. Siegel, B. Thielen, B. Benkert, M. Kucklick, M. Schobert, J. Barthelmes, C. Ebeling, I. Haddad, M. Scheer, A. Grote, K. Hiller, B. Bunk, K. Schreiber, I. Retter, D. Schomburg and D. Jahn (2007). "SYSTOMONAS -- an integrated database for systems biology analysis of *Pseudomonas*." *Nucleic Acids Research* **35**(suppl_1): D533-537.

Choi, E. N., M. C. Cho, Y. Kim, C.-K. Kim and K. Lee (2003). "Expansion of growth substrate range in *Pseudomonas putida* F1 by mutations in both *cymR* and todS, *which* recruit a ring-fission hydrolase CmtE and induce the *tod* catabolic operon, respectively." *Microbiology* **149**(3): 795-805.

Chou, C.-H., W.-C. Chang, C.-M. Chiu, C.-C. Huang and H.-D. Huang (2009). "FMM: a web server for metabolic pathway reconstruction and comparative analysis." *Nucleic Acids Research* **37**(suppl_2): W129-134.

Chou, H.-H., A.-P. Hsia, D. L. Mooney and P. S. Schnable (2004). "Picky: oligo microarray design for large genomes." *Bioinformatics* **20**(17): 2893-2902.

Chung, W.-H., S.-K. Rhee, X.-F. Wan, J.-W. Bae, Z.-X. Quan and Y.-H. Park (2005). "Design of long oligonucleotide probes for functional gene detection in a microbial community." *Bioinformatics* **21**(22): 4092-4100.

Clemente, A. R., T. A. Anazawa and L. R. Durrant (2001). "Biodegradation of polycylic aromatic hydrocarbons by soil fungi." *Brazilian Journal of Microbiology* **32**: 255-261.

Cochrane, G. R. and M. Y. Galperin (2010). "The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources." *Nucleic Acids Research* **38**(suppl_1): D1-4.

Conesa, A., S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." *Bioinformatics* **21**(18): 3674-3676.

Coppotelli, B., A. Ibarrolaza, M. Del Panno and I. Morelli (2008). "Effects of the Inoculant Strain *Sphingomonas paucimobilis* 20006FA on Soil Bacterial Community and Biodegradation in Phenanthrene-contaminated Soil." *Microbial Ecology* **55**(2): 173-183.

Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." *Nucleic Acids Research* **16**(22): 10881-10890.

Davies, J. and W. C. Evans (1964). "Oxidative Metabolism of Naphthalene by Soil Pseudomonads. The ring fission mechanism." *The Biochemical Journal* **91**(2): 251-261.

Davis, K. E. R., S. J. Joseph and P. H. Janssen (2005). "Effects of Growth Medium, Inoculum Size, and Incubation Time on Culturability and Isolation of Soil Bacteria." *Applied and Environmental Microbiology* **71**(2): 826-834.

Davison, J. (2002). "Towards safer vectors for the field release of recombinant bacteria." *Environmental Biosafety Research* **1**(19): 9-18.

Demaneche, S., C. Meyer, J. Micoud, M. Louwagie, J. C. Willison and Y. Jouanneau (2004). "Identification and Functional Analysis of Two Aromatic-Ring-Hydroxylating Dioxygenases from a *Sphingomonas* Strain That Degrades Various Polycyclic Aromatic Hydrocarbons." *Applied and Environmental Microbiology* **70**(11): 6714-6725.

Deng, Y., Z. He, J. Van Nostrand and J. Zhou (2008). "Design and analysis of mismatch probes for long oligonucleotide microarrays." *BMC Genomics* **9**(1): 491.

Dennis, P., E. A. Edwards, S. N. Liss and R. Fulthorpe (2003). "Monitoring Gene Expression in Mixed Microbial Communities by Using DNA Microarrays." *Applied and Environmental Microbiology* **69**(2): 769-778.

Deprince, A. (2003). La faune du sol : diversité, méthodes d'étude, fonctions et perspectives. Le Courrier de l'Environnement. INRA. **49**.

Desai, A., R. Autenrieth, P. Dimitriou-Christidis and T. McDonald (2008). "Biodegradation kinetics of select polycyclic aromatic hydrocarbon (PAH) mixtures by *Sphingomonas paucimobilis* EPA505." *Biodegradation* **19**(2): 223-233.

DeSantis, T., E. Brodie, J. Moberg, I. Zubieta, Y. Piceno and G. Andersen (2007). "High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment." *Microbial Ecology* **53**(3): 371-383.

Descorps-Declère, S., F. Lemoine, Q. Sculo, O. Lespinet and B. Labedan (2008). "The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species." *Biochimie* **90**(4): 595-608.

Dharmadi, Y. and R. Gonzalez (2004). "DNA Microarrays: Experimental Issues, Data Analysis, and Application to Bacterial Systems." *Biotechnology Progress* **20**(5): 1309-1324.

Dimitriou-Christidis, P. and R. Autenrieth (2007). "Kinetics of biodegradation of binary and ternary mixtures of PAHs." *Biotechnology and Bioengineering* **97**(4): 788-800.

Dinsdale, E. A., R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White and F. Rohwer (2008). "Functional metagenomic profiling of nine biomes." *Nature* **452**(7187): 629-632.

Duchaufour, P. (2001). **Introduction à la science du sol : Sol - Végétation - Environnement**, Dunod.

Dufayard, J.-F., L. Duret, S. Penel, M. Gouy, F. Rechenmann and G. Perrière (2005). "Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases." *Bioinformatics* **21**(11): 2596-2603.

Dufva, M. (2009a). Fabrication of DNA Microarray. DNA Microarrays for Biomedical Research**:** 63-79.

Dufva, M. (2009b). Introduction to Microarray Technology. DNA Microarrays for Biomedical Research**:** 1-22.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research* **32**(5): 1792-1797.

Ehlers, G. A. C. and A. P. Loibner (2006). "Linking organic pollutant (bio)availability with geosorbent properties and biomimetic methodology: A review of geosorbent characterisation and (bio)availability prediction." *Environmental Pollution* **141**(3): 494-512.

Ehrenreich, A. (2006). "DNA microarray technology for the microbiologist: an overview." *Applied Microbiology and Biotechnology* **73**(2): 255-273.

Eklund, A. C., P. Friis, R. Wernersson and Z. Szallasi (2009). "Optimization of the BLASTN substitution matrix for prediction of non-specific DNA microarray hybridization." *Nucleic Acids Research*: gkp1116.

Eldor A., P. (2007). **Soil microbiology, ecology, and biochemistry**. Amsterdam, Elsevier Science & Technology Books.

Elliott, B., M. Kirac, A. Cakmak, G. Yavas, S. Mayes, E. Cheng, Y. Wang, C. Gupta, G. Ozsoyoglu and Z. Meral Ozsoyoglu (2008). "PathCase: pathways database system." *Bioinformatics* **24**(21): 2526-2533.

Emrich, S. J., M. Lowe and A. L. Delcher (2003). "PROBEmer: a web-based software tool for selecting optimal DNA oligos." *Nucleic Acids Research* **31**(13): 3746-3750.

Felsenstein, J. (1978). "Cases in which parsimony and compatibility methods will be positively misleading." *Systematic Zoology* **27**(4): 401-410.

Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." *Journal of Molecular Evolution* **17**(6): 368-376.

Fernández-Luqueño, F., C. Valenzuela-Encinas, R. Marsch, C. Martínez-Suárez, E. Vázquez-Núñez and L. Dendooven (2010). "Microbial communities to mitigate contamination of PAHs in soil—possibilities and challenges: a review." *Environmental Science and Pollution Research* **doi: 10.1007/s11356-010-0371-6**.

Fernley, H. N. and W. C. Evans (1958). "Oxidative metabolism of polycyclic hydrocarbons by soil Pseudomonads." *Nature* **182**(4632): 373-375.

Ferrari, B. C., S. J. Binnerup and M. Gillings (2005). "Microcolony Cultivation on a Soil Substrate Membrane System Selects for Previously Uncultured Soil Bacteria." *Applied and Environmental Microbiology* **71**(12): 8714-8720.

Ferraro, D. J., L. Gakhar and S. Ramaswam (2005). "Rieske business: Structure–function of Rieske non-heme oxygenases." *Biochemical and Biophysical Research Communications* **338**: 175-190.

Fierer, N., M. Bradford and R. Jackson (2007a). "Toward an ecological classification of soil bacteria." *Ecology* **88**(6): 1354-1364.

Fierer, N., M. Breitbart, J. Nulton, P. Salamon, C. Lozupone, R. Jones, M. Robeson, R. A. Edwards, B. Felts, S. Rayhawk, R. Knight, F. Rohwer and R. B. Jackson (2007b). "Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil." *Applied and Environmental Microbiology* **73**(21): 7059-7066.

Fink, W. L. (1986). "Microcomputers and Phylogenetic Analysis." *Science* **234**(4780): 1135-1139.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy and A. Bateman (2010). "The Pfam protein families database." *Nucleic Acids Research* **38**(suppl_1): D211-222.

Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." *Science* **155**(760): 279-284.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and e. al (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." *Science* **269**(5223): 496-512.

Fontaine, S., S. Barot, P. Barre, N. Bdioui, B. Mary and C. Rumpel (2007). "Stability of organic carbon in deep soil layers controlled by fresh carbon supply." *Nature* **450**(7167): 277-280.

Fuchedzhieva, N., D. Karakashev and I. Angelidaki (2008). "Anaerobic biodegradation of fluoranthene under methanogenic conditions in presence of surface-active compounds." *Journal of Hazardous Materials* **153**(1-2): 123-127.

Fulton, D., Y. Li, M. Laird, B. Horsman, F. Roche and F. Brinkman (2006). "Improving the specificity of high-throughput ortholog prediction." *BMC Bioinformatics* **7**(1): 270.

Galvão, T., W. Mohn and V. de Lorenzo (2005). "Exploring the microbial biodegradation and biotransformation gene pool." *Trends in Biotechnology* **23**(10): 497-506.

Gan, S., E. V. Lau and H. K. Ng (2009). "Remediation of soils contaminated with polycyclic aromatic hydrocarbons (PAHs)." *Journal of Hazardous Materials* **172**(2-3): 532-549.

Gao, H., Z. K. Yang, T. J. Gentry, L. Wu, C. W. Schadt and J. Zhou (2007). "Microarray-Based Analysis of Microbial Community RNAs by Whole-Community RNA Amplification." *Applied and Environmental Microbiology* **73**(2): 563-571.

Gao, J., L. B. M. Ellis and L. P. Wackett (2010). "The University of Minnesota Biocatalysis/Biodegradation Database: improving public access." *Nucleic Acids Research* **38**(suppl_1): D488-491.

Gaspar, M., M. Cabello, M. Cazau and R. Pollero (2002). "Effect of phenanthrene and *Rhodotorula glutinis* on arbuscular mycorrhizal fungus colonization of maize roots." *Mycorrhiza* **12**(2): 55-59.

Gentry, T., G. Wickham, C. Schadt, Z. He and J. Zhou (2006). "Microarray Applications in Microbial Ecology Research." *Microbial Ecology* **52**(2): 159-175.

George, C. E., G. R. Lightsey, I. Jun and J. Fan (1992). "Soil decontamination via microwave and radio frequency co-volatilization." *Environmental Progress* **11**(3): 216-219.

Gerlt, J. and P. Babbitt (2000). "Can sequence determine function?" *Genome Biology* **1**(5): reviews0005.1–reviews0005.10.

Goesmann, A., M. Haubrock, F. Meyer, J. Kalinowski and R. Giegerich (2002). "PathFinder: reconstruction and dynamic visualization of metabolic pathways." *Bioinformatics* **18**(1): 124-129.

Goffard, N. and G. Weiller (2007). "PathExpress: a web-based tool to identify relevant pathways in gene expression data." *Nucleic Acids Research* **35**(suppl_2): W176-181.

Gordon, P. M. K. and C. W. Sensen (2004). "Osprey: a comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays." *Nucleic Acids Research* **32**(17): e133.

Gough, J., K. Karplus, R. Hughey and C. Chothia (2001). "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure." *Journal of Molecular Biology* **313**(4): 903-919.

Goyal, A. K. and G. J. Zylstra (1997). "Genetics of naphthalene and phenanthrene degradation by *Comamonas testosteroni*." *Journal of Industrial Microbiology and Biotechnology* **19**: 401-407.

Granjeaud, S., F. Bertucci and B. R. Jordan (1999). "Expression profiling: DNA arrays in many guises." *Bioessays* **21**(9): 781-790.

Grassot, J., G. Mouchiroud and G. Perrière (2003). "RTKdb: database of receptor tyrosine kinase." *Nucleic Acids Research* **31**(1): 353-358.

Gresham, D., B. Curry, A. Ward, D. B. Gordon, L. Brizuela, L. Kruglyak and D. Botstein (2010). "Optimized detection of sequence variation in heterozygous genomes using DNA microarrays with isothermal-melting probes." *Proceedings of the National Academy of Sciences of the United States of America* **107**(4): 1482-1487.

Gresham, D., M. J. Dunham and D. Botstein (2008). "Comparing whole genomes using DNA microarrays." *Nature Reviews Genetics* **9**: 291-302.

Gromiha, M. M., Y. Yabuki, M. X. Suresh, A. M. Thangakani, M. Suwa and K. Fukui (2009). "TMFunction: database for functional residues in membrane proteins." *Nucleic Acids Research* **37**(suppl_1): D201-204.

Grossetete, S., B. Labedan and O. Lespinet (2010). "FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology." *BMC Genomics* **11**(1): 81.

Guschin, D. Y., B. K. Mobarry, D. Proudnikov, D. A. Stahl, B. E. Rittmann and A. D. Mirzabekov (1997). "Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology." *Applied and Environmental Microbiology* **63**(6): 2397-2402.

Habe, H. and T. Omori (2003). "Genetics of Polycyclic Aromatic Hydrocarbon Metabolism in Diverse Aerobic Bacteria." *Bioscience, Biotechnology, and Biochemistry* **67**(2): 225-243.

Haft, D. H., J. D. Selengut and O. White (2003). "The TIGRFAMs database of protein families." *Nucleic Acids Research* **31**(1): 371-373.

Hall, B. G., A. Pikis and J. Thompson (2009). "Evolution and Biochemistry of Family 4 Glycosidases: Implications for Assigning Enzyme Function in Sequence Annotations." *Molecular Biology and Evolution* **26**(11): 2487-2497.

Hall, N. (2007). "Advanced sequencing technologies and their wider impact in microbiology." *Journal of Experimental Biology* **210**(9): 1518-1525.

Halmemies, S., S. Gröndahl, M. Arffman, K. Nenonen and T. Tuhkanen (2003). "Vacuum extraction based response equipment for recovery of fresh fuel spills from soil." *Journal of Hazardous Materials* **97**(1-3): 127-143.

Hammer, Ø., D. A. T. Harper and P. D. Ryan (2001). "PAST: Paleontological Statistics Software Package for Education and Data Analysis." *Palaeontologia Electronica* **4**(1): 9pp.

Harayama, S., M. Kok and E. L. Neidle (1992). "Functional and Evolutionary Relationships Among Diverse Oxygenases." *Annual Review of Microbiology* **46**(1): 565-601.

Haritash, A. K. and C. P. Kaushik (2009). "Biodegradation aspects of Polycyclic Aromatic Hydrocarbons (PAHs): A review." *Journal of Hazardous Materials* **169**(1-3): 1-15.

He, Z., T. J. Gentry, C. W. Schadt, L. Wu, J. Liebich, S. C. Chong, Z. Huang, W. Wu, B. Gu, P. Jardine, C. Criddle and J. Zhou (2007). "GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes." *The ISME Journal* **1**(1): 67-77.

He, Z., L. Wu, X. Li, M. W. Fields and J. Zhou (2005). "Empirical Establishment of Oligonucleotide Probe Design Criteria." *Applied and Environmental Microbiology* **71**(7): 3753-3760.

He, Z. and J. Zhou (2008). "Empirical Evaluation of a New Method for Calculating Signal-to-Noise Ratio for Microarray Data Analysis." *Applied and Environmental Microbiology* **74**(10): 2957-2966.

Hearn, E. M., J. J. Dennis, M. R. Gray and J. M. Foght (2003). "Identification and Characterization of the *emhABC* Efflux System for Polycyclic Aromatic Hydrocarbons in *Pseudomonas fluorescens* cLP6a." *Journal of Bacteriology* **185**(21): 6233-6240.

Hemme, C. L., Y. Deng, T. J. Gentry, M. W. Fields, L. Wu, S. Barua, K. Barry, S. G. Tringe, D. B. Watson, Z. He, T. C. Hazen, J. M. Tiedje, E. M. Rubin and J. Zhou (2010). "Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community." *The ISME Journal* **4**(5): 660-672.

Henikoff, S. and J. Henikoff (1993). "Performance evaluation of amino acid substitution matrices." *Proteins* **17**(1): 49-61.

Hernandez-Raquet, G., H. Budzinski, P. Caumette, P. Dabert, K. Le Ménach, G. Muyzer and R. Duran (2006). "Molecular diversity studies of bacterial communities of oil polluted microbial mats from the Etang de Berre (France)." *FEMS Microbiology Ecology* **58**(3): 550-562.

Hertzsch, J.-M., R. Sturman and S. Wiggins (2007). "DNA Microarrays: Design Principles for Maximizing Ergodic, Chaotic Mixing." *Small* **3**(2): 202-218.

Huang, H., C. Xiao and C. H. Wu (2000). "ProClass protein family database." *Nucleic Acids Research* **28**(1): 273-276.

Huang, W., P. Peng, Z. Yu and J. Fu (2003). "Effects of organic matter heterogeneity on sorption and desorption of organic contaminants by soils and sediments." *Applied Geochemistry* **18**: 955-972.

Huffman, J. L. and R. G. Brennan (2002). "Prokaryotic transcription regulators: more than just the helix-turn-helix motif." *Current Opinion in Structural Biology* **12**(1): 98-106.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats (2009). "InterPro: the integrative protein signature database." *Nucleic Acids Research* **37**(suppl_1): D211-215.

Isken, S. and J. A. de Bont (1998). "Bacteria tolerant to organic solvents." *Extremophiles* **2**(3): 229-238.

Iwabuchi, T. and S. Harayama (1998a). "Biochemical and Genetic Characterization of trans-2'-Carboxybenzalpyruvate Hydratase-Aldolase from a Phenanthrene-Degrading *Nocardioides* Strain." *Journal of Bacteriology* **180**(4): 945-949.

Iwabuchi, T. and S. Harayama (1998b). "Biochemical and Molecular Characterization of 1-Hydroxy-2-naphthoate Dioxygenase from *Nocardioides* sp. KP7." *The Journal of Biological Chemistry* **273**(14): 8332-8336.

Iwai, S., F. Kurisu, H. Urakawa, O. Yagi and H. Furumai (2007). "Development of a 60-mer oligonucleotide microarray on the basis of benzene monooxygenase gene diversity." *Applied Microbiology and Biotechnology* **75**(4): 929-939.

Iwai, S., F. Kurisu, H. Urakawa, O. Yagi and H. Furumai (2010). "Characterization of monooxygenase gene diversity in benzene-amended soils." *Letters in Applied Microbiology* **50**(2): 138-145.

Iwai, S., F. Kurisu, H. Urakawa, O. Yagi, I. Kasuga and H. Furumai (2008). "Development of an oligonucleotide microarray to detect di- and monooxygenase genes for benzene degradation in soil." *FEMS Microbiology Letters* **285**(1): 111-121.

Jabado, O. J., G. Palacios, V. Kapoor, J. Hui, N. Renwick, J. Zhai, T. Briese and W. I. Lipkin (2006). "Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments." *Nucleic Acids Research* **34**(22): 6605-6611.

Jakoncic, J., Y. Jouanneau, C. Meyer and V. Stojanoff (2007a). "The catalytic pocket of the ring-hydroxylating dioxygenase from *Sphingomonas* CHY-1." *Biochemical and Biophysical Research Communications* **352**(4): 861–866.

Jakoncic, J., Y. Jouanneau, C. Meyer and V. Stojanoff (2007b). "The crystal structure of the ring-hydroxylating dioxygenase from *Sphingomonas* CHY-1." *The FEBS Journal* **274**(10): 2470-2481.

Jeon, C. O., M. Park, H.-S. Ro, W. Park and E. L. Madsen (2006). "The Naphthalene Catabolic (*nag*) Genes of *Polaromonas naphthalenivorans* CJ2: Evolutionary Implications for Two Gene Clusters and Novel Regulatory Control." *Applied and Environmental Microbiology* **72**(2): 1086-1095.

Johnsen, A. R. and U. Karlson (2004). "Evaluation of bacterial strategies to promote the bioavailability of polycyclic aromatic hydrocarbons." *Applied Microbiology and Biotechnology* **63**(4): 452-459.

Jones, R. M., B. Britt-Compton and P. A. Williams (2003). "The Naphthalene Catabolic (*nag*) Genes of *Ralstonia* sp. Strain U2 Are an Operon That Is Regulated by NagR, a LysR-Type Transcriptional Regulator." *Journal of Bacteriology* **185**(19): 5847-5853.

Joux, J., J. C. Bertrand, R. De Wit, V. Grossi, L. Intertaglia, P. Lebaron, V. Michotey, P. Normand, P. Peyret, P. Raimbault, C. Tamburini and L. Urios (2010). **Les biopuces en Ecologie Microbienne (In : Méthodes d'études des micro-organismes dans l'environnement)**.

Kaderali, L. and A. Schliep (2002). "Selecting signature oligonucleotides to identify organisms using DNA arrays." *Bioinformatics* **18**(10): 1340-1349.

Kallimanis, A., S. Frillingos, C. Drainas and A. Koukkou (2007). "Taxonomic identification, phenanthrene uptake activity, and membrane lipid alterations of the PAH degrading *Arthrobacter* sp. strain Sphe3." *Applied Microbiology and Biotechnology* **76**(3): 709-717.

Kan, A. T., G. Fu and M. B. Tomson (1994). "Adsorption/Desorption Hysteresis in Organic Pollutant and Soil/Sediment Interaction." *Environmental Science & Technology* **28**(5): 859-867.

Kane, M. D., T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas and S. J. Madore (2000). "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays." *Nucleic Acids Research* **28**(22): 4552-4557.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi (2008). "KEGG for

linking genomes to life and the environment." *Nucleic Acids Research* **36**(suppl_1): D480-484.

Kaplan, N., O. Sasson, U. Inbar, M. Friedlich, M. Fromer, H. Fleischer, E. Portugaly, N. Linial and M. Linial (2005). "ProtoNet 4.0: A hierarchical classification of one million protein sequences." *Nucleic Acids Research* **33**(suppl_1): D216-218.

Karimpour-Fard, A., S. Leach, R. Gill and L. Hunter (2008). "Predicting protein linkages in bacteria: Which method is best depends on task." *BMC Bioinformatics* **9**(1): 397.

Karp, P. D., S. Paley and P. Romero (2002). "The Pathway Tools software." *Bioinformatics* **18**(suppl_1): S225-232.

Keck, A., D. Conradt, A. Mahler, A. Stolz, R. Mattes and J. Klein (2006). "Identification and functional analysis of the genes for naphthalenesulfonate catabolism by *Sphingomonas xenophaga* BN6." *Microbiology* **152**(7): 1929-1940.

Kelley, I., J. P. Freeman and C. E. Cerniglia (1990). "Identification of metabolites from degradation of naphthalene by a *Mycobacterium* sp." *Biodegradation* **1**(4): 283-290.

Keseler, I. M., C. Bonavides-Martinez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson, M. Krummenacker, L. M. Nolan, S. Paley, I. T. Paulsen, M. Peralta-Gil, A. Santos-Zavaleta, A. G. Shearer and P. D. Karp (2009). "EcoCyc: A comprehensive view of *Escherichia coli* biology." *Nucleic Acids Research* **37**(suppl_1): D464-470.

Khalladi, R., O. Benhabiles, F. Bentahar and N. Moulai-Mostefa (2009). "Surfactant remediation of diesel fuel polluted soil." *Journal of Hazardous Materials* **164**(2-3): 1179-1184.

Khan, A. A., R.-F. Wang, W.-W. Cao, D. R. Doerge, D. Wennerstrom and C. E. Cerniglia (2001). "Molecular Cloning, Nucleotide Sequence, and Expression of Genes Encoding a Polycyclic Aromatic Ring Dioxygenase from Mycobacterium sp. Strain PYR-1." *Applied and Environmental Microbiology* **67**(8): 3577-3585.

Khan, F. I., T. Husain and R. Hejazi (2004). "An overview and analysis of site remediation technologies." *Journal of Environmental Management* **71**(2): 95-122.

Kikuchi, Y., Y. Yasukochi, Y. Nagata, M. Fukuda and M. Takagi (1994). "Nucleotide sequence and functional analysis of the *meta*-cleavage pathway involved in biphenyl and polychlorinated biphenyl degradation in *Pseudomonas* sp. strain KKS102." *Journal of Bacteriology* **176**(14): 4269-4276.

Kim, D., Y.-S. Kim, S.-K. Kim, S. W. Kim, G. J. Zylstra, Y. M. Kim and E. Kim (2002). "Monocyclic Aromatic Hydrocarbon Degradation by *Rhodococcus* sp. Strain DK17." *Applied and Environmental Microbiology* **68**(7): 3270-3278.

Kim, E., G. Zylstra, J. Freeman, T. Heinze, J. Deck and C. Cerniglia (1997). "Evidence for the role of 2-hydroxychromene-2-carboxylate isomerase in the degradation of anthracene by *Sphingomonas yanoikuyae* B1." *FEMS Microbiology Letters* **153**(2): 479-484.

Kim, E. and G. J. Zylstra (1999). "Functional analysis of genes involved in biphenyl, naphthalene, phenanthrene, and m -xylene degradation by *Sphingomonas yanoikuyae* B1." *Journal of Industrial Microbiology and Biotechnology* **23**(4): 294-302.

Kim, J.-S. and D. E. Crowley (2007a). "Microbial Diversity in Natural Asphalts of the Rancho La Brea Tar Pits." *AEM* **73**(14): 4579.

Kim, J.-S. and D. E. Crowley (2007b). "Microbial Diversity in Natural Asphalts of the Rancho La Brea Tar Pits." *Applied and Environmental Microbiology* **73**(14): 4579.

Kim, M., S. Bae, M. Seol, J.-H. Lee and Y.-S. Oh (2008). "Monitoring nutrient impact on bacterial community composition during bioremediation of anoxic PAH-contaminated sediment." *Journal of Microbiology* **46**(6): 615-623.

Kim, S.-J., O. Kweon, J. P. Freeman, R. C. Jones, M. D. Adjei, J.-W. Jhoo, R. D. Edmondson and C. E. Cerniglia (2006). "Molecular Cloning and Expression of Genes Encoding a Novel Dioxygenase Involved in Low- and High-Molecular-Weight Polycyclic Aromatic Hydrocarbon Degradation in *Mycobacterium vanbaalenii* PYR-1." *Applied and Environmental Microbiology* **72**(2): 1045-1054.

Kim, S.-J., O. Kweon, R. C. Jones, J. P. Freeman, R. D. Edmondson and C. E. Cerniglia (2007). "Complete and Integrated Pyrene Degradation Pathway in *Mycobacterium vanbaalenii* PYR-1 Based on Systems Biology." *Journal of Bacteriology* **189**(2): 464-472.

Kiyohara, H. and K. Nagao (1978). "The Catabolism of Phenanthrene and Naphthalene by Bacteria." *Journal of General Microbiology* **105**(1): 69-75.

Klukas, C. and F. Schreiber (2007). "Dynamic exploration and editing of KEGG pathway diagrams." *Bioinformatics* **23**(3): 344-350.

Krause, A., J. Stoye and M. Vingron (2000). "The SYSTERS protein sequence cluster set." *Nucleic Acids Research* **28**(1): 270-272.

Kreil, D. P., R. R. Russell, S. Russell and a. B. O. Alan Kimmel (2006). Microarray Oligonucleotide Probes. Methods in Enzymology, Academic Press. **Volume 410:** 73-98.

Kriventseva, E. V., W. Fleischmann, E. M. Zdobnov and R. Apweiler (2001). "CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins." *Nucleic Acids Research* **29**(1): 33-36.

Kriventseva, E. V., N. Rahman, O. Espinosa and E. M. Zdobnov (2008). "OrthoDB: the hierarchical catalog of eukaryotic orthologs." *Nucleic Acids Research* **36**(suppl_1): D271-275.

Krivobok, S., S. Kuony, C. Meyer, M. Louwagie, J. C. Willison and Y. Jouanneau (2003). "Identification of Pyrene-Induced Proteins in *Mycobacterium* sp. Strain 6PY1: Evidence for Two Ring-Hydroxylating Dioxygenases." *Journal of Bacteriology* **185**(13): 3828-3841.

Kulakov, L. A., C. C. Allen, D. A. Lipscomb and M. J. Larkin (2000). "Cloning and characterization of a novel cis-naphthalene dihydrodiol dehydrogenase gene (*narB*) from *Rhodococcus* sp. NCIMB12038." *FEMS Microbiology Letters* **182**(2): 327-331.

Kulikova, T., R. Akhtar, P. Aldebert, N. Althorpe, M. Andersson, A. Baldwin, K. Bates, S. Bhattacharyya, L. Bower, P. Browne, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, G. Hoad, C. Kanz, C. Lee, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, F. Nardone, M. P. G. Pastor, S. Plaister, S. Sobhany, P. Stoehr, R. Vaughan, D. Wu, W. Zhu and R. Apweiler (2007). "EMBL Nucleotide Sequence Database in 2006." *Nucleic Acids Research* **35**(suppl_1): D16-20.

Kumar, M. and S. Khanna (2010). "Diversity of 16S rRNA and dioxygenase genes detected in coal-tar-contaminated site undergoing active bioremediation." *Journal of Applied Microbiology* **108**(4): 1252-1262.

Kuzniar, A., R. C. H. J. van Ham, S. Pongor and J. A. M. Leunissen (2008). "The quest for orthologs: finding the corresponding gene across genomes." *Trends in Genetics* **24**(11): 539.

Kweon, O., S.-J. Kim, S. Baek, J.-C. Chae, M. Adjei, D.-H. Baek, Y.-C. Kim and C. Cerniglia (2008). "A new classification system for bacterial Rieske non-heme iron aromatic ring-hydroxylating oxygenases." *BMC Biochemistry* **9**(1): 11-22.

Kweon, O., S.-J. Kim, R. C. Jones, J. P. Freeman, M. D. Adjei, R. D. Edmondson and C. E. Cerniglia (2007). "A Polyomic Approach To Elucidate the Fluoranthene-

Degradative Pathway in *Mycobacterium vanbaalenii* PYR-1." *Journal of Bacteriology* **189**(13): 4635-4647.

Lane, D. J. (1991). 16S/23S rRNA sequencing. Nucleic acid techniques in bacterial systematics. e. E. Stackebrandt and M. Goodfellow. New York, NY, John Wiley and Sons**:** 115-175.

Larkin, M. J., C. C. R. Allen, L. A. Kulakov and D. A. Lipscomb (1999). "Purification and Characterization of a Novel Naphthalene Dioxygenase from *Rhodococcus* sp. Strain NCIMB12038." *Journal of Bacteriology* **181**(19): 6200-6204.

Laurie, A. D. and G. Lloyd-Jones (1999). "The *phn* Genes of *Burkholderia* sp. Strain RP007 Constitute a Divergent Gene Cluster for Polycyclic Aromatic Hydrocarbon Catabolism." *Journal of Bacteriology* **181**(2): 531-540.

Le Fèvre, F., S. Smidtas and V. Schächter (2007). "Cyclone: java-based querying and computing with Pathway/Genome databases." *Bioinformatics* **23**(10): 1299-1300.

Leadbetter, J. (2003). "Cultivation of recalcitrant microbes: cells are alive, well and revealing their secrets in the 21st century laboratory." *Current Opinion in Microbiology* **6**(3): 274-281.

Lebkowska, M., E. Karwowska and E. Miaśkiewicz (1995). "Isolation and identification of bacteria from petroleum derivatives contaminated soil." *Acta Microbiologica Polonica* **44**(3-4): 297-303.

Lee, S.-E., J.-S. Seo, Y.-S. Keum, J. H. Lee and C.-H. Li (2007). "Fluoranthene metabolism and associated proteins in *Mycobacterium* sp. JS14." *Proteomics* **7**(12): 2059-2069.

Leglize, P., S. Alain, B. Jacques and L. Corinne (2008). "Adsorption of phenanthrene on activated carbon increases mineralization rate by specific bacteria." *Journal of Hazardous Materials* **151**(2-3): 339-347.

Lei, A.-P., Z.-L. Hu, Y.-S. Wong and N. F.-Y. Tam (2007). "Removal of fluoranthene and pyrene by different microalgal species." *Bioresource Technology* **98**(2): 273-280.

Lemoine, F., B. Labedan and O. Lespinet (2008). "SynteBase/SynteView: a tool to visualize gene order conservation in prokaryotic genomes." *BMC Bioinformatics* **9**(1): 536.

Lemoine, F., O. Lespinet and B. Labedan (2007). "Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data." *BMC Evolutionary Biology* **7**(1): 237.

Lemoine, S., F. Combes and S. Le Crom (2009). "An evaluation of custom microarray applications: the oligonucleotide design challenge." *Nucleic Acids Research* **37**(6): 1726-1739.

Leparc, G. G., T. Tuchler, G. Striedner, K. Bayer, P. Sykacek, I. L. Hofacker and D. P. Kreil (2009). "Model-based probe set optimization for high-performance microarrays." *Nucleic Acids Research* **37**(3): e18.

Letunic, I., T. Doerks and P. Bork (2009). "SMART 6: recent updates and new developments." *Nucleic Acids Research* **37**(suppl_1): D229-232.

Leys, N. M. E. J., A. Ryngaert, L. Bastiaens, W. Verstraete, E. M. Top and D. Springael (2004). "Occurrence and Phylogenetic Diversity of Sphingomonas Strains in Soils Contaminated with Polycyclic Aromatic Hydrocarbons." *AEM* **70**(4): 1944-1955.

Li, C.-H., H.-W. Zhou, Y.-S. Wong and N. F.-Y. Tam (2009a). "Vertical distribution and anaerobic biodegradation of polycyclic aromatic hydrocarbons in mangrove sediments in Hong Kong, South China." *Science of The Total Environment* **407**(21): 5772-5779.

Li, C., X. Li, Y. Miao, Q. Wang, W. Jiang, C. Xu, J. Li, J. Han, F. Zhang, B. Gong and L. Xu (2009b). "SubpathwayMiner: a software package for flexible identification of pathways." *Nucleic Acids Research* **37**(19): e131.

Li, F. and G. D. Stormo (2001). "Selection of optimal DNA oligos for gene expression arrays." *Bioinformatics* **17**(11): 1067-1076.

Li, G., D. Che and Y. Xu (2009c). "A universal operon predictor for prokaryotic genomes." *Journal of Bioinformatics and Computational Biology* **7**(1): 19-38.

Li, S., A. Pozhitkov and M. Brouwer (2008). "A competitive hybridization model predicts probe signal intensity on high density DNA microarrays." *Nucleic Acids Research* **36**(20): 6585-6591.

Li, X., Z. He and J. Zhou (2005). "Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation." *Nucleic Acids Research* **33**(19): 6114-6123.

Liang, Y., G. Li, J. D. Van Nostrand, Z. He, L. Wu, Y. Deng, X. Zhang and J. Zhou (2009a). "Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field." *FEMS Microbiology Ecology* **70**(2): 324-333.

Liang, Y., J. D. V. Nostrand, J. Wang, X. Zhang, J. Zhou and G. Li (2009b). "Microarray-based functional gene analysis of soil microbial communities during ozonation and biodegradation of crude oil." *Chemosphere* **75**(2): 193-199.

Liebich, J., C. W. Schadt, S. C. Chong, Z. He, S.-K. Rhee and J. Zhou (2006). "Improvement of Oligonucleotide Probe Design Criteria for Functional Gene Microarrays in Environmental Applications." *Applied and Environmental Microbiology* **72**(2): 1688-1691.

Lloyd-Jones, G. and P. C. Lau (1997). "Glutathione S-transferase-encoding gene as a potential probe for environmental bacterial isolates capable of degrading polycyclic aromatic hydrocarbons." *Applied and Environmental Microbiology* **63**(8): 3285-3290.

Loy, A., A. Lehner, N. Lee, J. Adamczyk, H. Meier, J. Ernst, K.-H. Schleifer and M. Wagner (2002). "Oligonucleotide Microarray for 16S rRNA Gene-Based Detection of All Recognized Lineages of Sulfate-Reducing Prokaryotes in the Environment." *Applied and Environmental Microbiology* **68**(10): 5064-5081.

Loy, A., C. Schulz, S. Lucker, A. Schopfer-Wendels, K. Stoecker, C. Baranyi, A. Lehner and M. Wagner (2005). "16S rRNA Gene-Based Oligonucleotide Microarray for Environmental Monitoring of the Betaproteobacterial Order "*Rhodocyclales*"." *Applied and Environmental Microbiology* **71**(3): 1373-1386.

Lozada, M., J. Riva Mercadal, L. Guerrero, W. Di Marzio, M. Ferrero and H. Dionisi (2008). "Novel aromatic ring-hydroxylating dioxygenase genes from coastal marine sediments of Patagonia." *BMC Microbiology* **8**(1): 50.

Lu, Z. J. and D. H. Mathews (2008). "OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics." *Nucleic Acids Research* **36**(suppl_2): W104-108.

Ludemann, A., D. Weicht, J. Selbig and J. Kopka (2004). "PaVESy: Pathway Visualization and Editing System." *Bioinformatics* **20**(16): 2841-2844.

MacLean, A. M., G. MacPherson, P. Aneja and T. M. Finan (2006). "Characterization of the {beta}-Ketoadipate Pathway in *Sinorhizobium meliloti*." *Applied and Environmental Microbiology* **72**(8): 5403-5413.

Mahadevan, P. and D. Seto (2010). "Rapid pair-wise synteny analysis of large bacterial genomes using web-based GeneOrder4.0." *BMC Research Notes* **3**(1): 41.

Makarova, K., A. Sorokin, P. Novichkov, Y. Wolf and E. Koonin (2007). "Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea." *Biology direct* **2**(1): 33.

Mallick, S., S. Chatterjee and T. K. Dutta (2007). "A novel degradation pathway in the assimilation of phenanthrene by *Staphylococcus* sp. strain PN/Y via meta-cleavage of 2-hydroxy-1-naphthoic acid: formation of trans-2,3-dioxo-5-(2'-hydroxyphenyl)-pent-4-enoic acid." *Microbiology* **153**(7): 2104-2115.

Maltsev, N., E. Glass, D. Sulakhe, A. Rodriguez, M. H. Syed, T. Bompada, Y. Zhang and M. D'Souza (2006). "PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways." *Nucleic Acids Research* **34**(suppl_1): D369-372.

Marchler-Bauer, A., J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang and S. H. Bryant (2009). "CDD: specific functional annotation with the Conserved Domain Database." *Nucleic Acids Research* **37**(suppl_1): D205-210.

Marcy, Y., P. Cousin, M. Rattier, G. Cerovic, G. Escalier, G. Béna, M. Guéron, L. McDonagh, F. le Boulaire, H. Bénisty, C. Weisbuch and J. Avarre (2008). "Innovative integrated system for real-time measurement of hybridization and melting on standard format microarrays." *Biotechniques* **44**(7): 913-920.

Margesin, R. and F. Schinner (2001). "Bioremediation (Natural Attenuation and Biostimulation) of Diesel-Oil-Contaminated Soil in an Alpine Glacier Skiing Area." *Applied and Environmental Microbiology* **67**(7): 3127-3133.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* **437**(7057): 376-380.

Marques, S., M.-T. Gallegos, M. Manzanera, A. Holtel, K. N. Timmis and J. L. Ramos (1998). "Activation and Repression of Transcription at the Double Tandem Divergent Promoters for the *xylR* and *xylS* Genes of the TOL Plasmid of *Pseudomonas putida*." *Journal of Bacteriology* **180**(11): 2889-2894.

Matveeva, O. V., D. H. Mathews, A. D. Tsodikov, S. A. Shabalina, R. F. Gesteland, J. F. Atkins and S. M. Freier (2003). "Thermodynamic criteria for high hit rate antisense oligonucleotide design." *Nucleic Acids Research* **31**(17): 4989-4994.

McDonald, A. G., K. F. Tipton and S. Boyce (2009). "Tracing metabolic pathways from enzyme data." *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1794**(9): 1364-1371.

McFall, S. M., S. A. Chugani and A. M. Chakrabarty (1998). "Transcriptional activation of the catechol and chlorocatechol operons: variations on a theme." *Gene* **223**(1-2): 257-267.

McQuain, M. K., K. Seale, J. Peek, T. S. Fisher, S. Levy, M. A. Stremler and F. R. Haselton (2004). "Chaotic mixer improves microarray hybridization." *Analytical Biochemistry* **325**(2): 215-226.

McShan, D. C., S. Rao and I. Shah (2003). "PathMiner: predicting metabolic pathways by heuristic search." *Bioinformatics* **19**(13): 1692-1698.

Meckenstock, R. U., M. Safinowski and C. Griebler (2004). "Anaerobic degradation of polycyclic aromatic hydrocarbons." *FEMS Microbiology Ecology* **49**(1): 27-36.

Mi, H., Q. Dong, A. Muruganujan, P. Gaudet, S. Lewis and P. D. Thomas (2010). "PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium." *Nucleic Acids Research* **38**(suppl_1): D204-210.

Mihelcic, J. R. and R. G. Luthy (1988a). "Degradation of polycyclic aromatic hydrocarbon compounds under various redox conditions in soil-water systems." *Applied and Environmental Microbiology* **54**(5): 1182-1187.

Mihelcic, J. R. and R. G. Luthy (1988b). "Microbial degradation of acenaphthene and naphthalene under denitrification conditions in soil-water systems." *Applied and Environmental Microbiology* **54**(5): 1188-1198.

Militon, C., S. Rimour, M. Missaoui, C. Biderre, V. Barra, D. Hill, A. Mone, G. Gagne, H. Meier, E. Peyretaillade and P. Peyret (2007). "PhylArray: phylogenetic probe design algorithm for microarray." *Bioinformatics* **23**(19): 2550-2557.

Mlecnik, B., M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo and Z. Trajanoski (2005). "PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways." *Nucleic Acids Research* **33**(suppl_2): W633-637.

Molina, M., N. González, L. Bautista, R. Sanz, R. Simarro, I. Sánchez and J. Sanz (2009). "Isolation and genetic identification of PAH degrading bacteria from a microbial consortium." *Biodegradation* **20**(6): 789-800.

Moody, J. D., J. P. Freeman, D. R. Doerge and C. E. Cerniglia (2001). "Degradation of Phenanthrene and Anthracene by Cell Suspensions of *Mycobacterium* sp. Strain PYR-1." *Applied and Environmental Microbiology* **67**(4): 1476-1483.

Morgenstern, B., K. Frech, A. Dress and T. Werner (1998). "DIALIGN: finding local similarities by multiple sequence alignment." *Bioinformatics* **14**(3): 290-294.

Mueckstein, U., G. Leparc, A. Posekany, I. Hofacker and D. Kreil (2010). "Hybridization thermodynamics of NimbleGen Microarrays." *BMC Bioinformatics* **11**(1): 35.

Mueller, J. G., P. J. Chapman, B. O. Blattmann and P. H. Pritchard (1990). "Isolation and characterization of a fluoranthene-utilizing strain of *Pseudomonas paucimobilis*." *Applied and Environmental Microbiology* **56**(4): 1079-1086.

Mulligan, C. N. and R. N. Yong (2004). "Natural attenuation of contaminated soils." *Environment International* **30**(4): 587-601.

Mulligan, C. N., R. N. Yong and B. F. Gibbs (2001). "Heavy metal removal from sediments by biosurfactants." *Journal of Hazardous Materials* **85**(1-2): 111-125.

Muratova, A., N. Pozdnyakova, S. Golubev, L. Wittenmayer, O. Makarov, W. Merbach and O. Turkovskaya (2009). "Oxidoreductase activity of sorghum root exudates in a phenanthrene-contaminated environment." *Chemosphere* **74**(8): 1031-1036.

Musy, A. and M. Soutter (1993). **Physique du sol**, Presses Polytechniques et Universitaires Romandes.

Nam, J. W., H. Nojiri, T. Yoshida, H. Habe, H. Yamane and T. Omori (2001). "New classification system for oxygenase components involved in ring-hydroxylating oxygenations." *Bioscience, Biotechnology, and Biochemistry* **65**(2): 254 - 263.

Navarro, R. R., H. Ichikawa, K. Morimoto and K. Tatsumi (2009). "Enhancing the release and plant uptake of PAHs with a water-soluble purine alkaloid." *Chemosphere* **76**(8): 1109-1113.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology* **48**(3): 443-453.

Neufeld, J. D., M. Wagner and J. C. Murrell (2007). "Who eats what, where and when? Isotope-labelling experiments are coming of age." *The ISME Journal* **1**(2): 103-110.

Ng, S.-K., Z. Zhang, S.-H. Tan and K. Lin (2003). "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes." *Nucleic Acids Research* **31**(1): 251-254.

Ní Chadhain, S., E. Moritz, E. Kim and G. Zylstra (2007). "Identification, cloning, and characterization of a multicomponent biphenyl dioxygenase from *Sphingobium yanoikuyae* B1." *Journal of Industrial Microbiology and Biotechnology* **34**(9): 605-613.

Nolan, T., R. Hands and S. Bustin (2006). "Quantification of mRNA using real-time RT-PCR." *Nature Protocols* **1**(3): 1559-1582.

Nordberg, E. K. (2005). "YODA: selecting signature oligonucleotides." *Bioinformatics* **21**(8): 1365-1370.

O'Brien, H. E., J. L. Parrent, J. A. Jackson, J.-M. Moncalvo and R. Vilgalys (2005). "Fungal Community Analysis by Large-Scale Sequencing of Environmental Samples." *Applied and Environmental Microbiology* **71**(9): 5544-5550.

O'Mahony, M. M., A. D. W. Dobson, J. D. Barnes and I. Singleton (2006). "The use of ozone in the remediation of polycyclic aromatic hydrocarbon contaminated soil." *Chemosphere* **63**(2): 307-314.

Oehm, S., D. Gilbert, A. Tauch, J. Stoye and A. Goesmann (2008). "Comparative Pathway Analyzer--a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms." *Nucleic Acids Research* **36**(suppl_2): W433-437.

Overbeek, R., N. Larsen, G. D. Pusch, M. D'Souza, E. S. Jr, N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov (2000). "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction." *Nucleic Acids Research* **28**(1): 123-125.

Park, J.-H., Y. Feng, P. Ji, T. C. Voice and S. A. Boyd (2003). "Assessment of Bioavailability of Soil-Sorbed Atrazine." *Applied and Environmental Microbiology* **69**(6): 3288-3298.

Park, S.-W., J.-Y. Lee, J.-S. Yang, K.-J. Kim and K. Baek (2009). "Electrokinetic remediation of contaminated soil with waste-lubricant oils and zinc." *Journal of Hazardous Materials* **169**(1-3): 1168-1172.

Parke, D. (1993). "Positive regulation of phenolic catabolism in *Agrobacterium tumefaciens* by the *pcaQ* gene in response to beta-carboxy-cis,cis-muconate." *Journal of Bacteriology* **175**(11): 3529-3535.

Parke, D. (1995). "Supraoperonic clustering of *pca* genes for catabolism of the phenolic compound protocatechuate in *Agrobacterium tumefaciens*." *Journal of Bacteriology* **177**(13): 3808-3817.

Pavlopoulos, G., A.-L. Wegener and R. Schneider (2008). "A survey of visualization tools for biological network analysis." *Biodata Mining* **1**(1): 12.

Pazos, F., D. Guijas, A. Valencia and V. De Lorenzo (2005). "MetaRouter: bioinformatics for bioremediation." *Nucleic Acids Research* **33**(suppl_1): D588-592.

Pearson, W. and D. Lipman (1988). "Improved tools for biological sequence comparison." *Proceedings of the National Academy of Sciences of the United States of America* **85**(8): 2444-2448.

Peng, R.-H., A.-S. Xiong, Y. Xue, X.-Y. Fu, F. Gao, W. Zhao, Y.-S. Tian and Q.-H. Yao (2008). "Microbial biodegradation of polyaromatic hydrocarbons." *FEMS Microbiology Reviews* **32**(6): 927-955.

Perrière, G. and C. Brochier-Armanet (2009). **Concepts et Méthodes en Phylogénie Moléculaire**. New York, Springer-Verlag, LLC.

Perrière, G., L. Duret and M. Gouy (2000). "HOBACGEN: Database System for Comparative Genomics in Bacteria." *Genome Research* **10**(3): 379-385.

Phale, P. S., A. Basu, P. D. Majhi, J. Deveryshetty, C. Vamsee-Krishna and R. Shrivastava (2007). "Metabolic Diversity in Bacterial Degradation of Aromatic Compounds." *OMICS: A Journal of Integrative Biology* **11**(3): 252-279.

Pietrokovski, S., J. G. Henikoff and S. Henikoff (1996). "The Blocks database--a system for protein classification." *Nucleic Acids Research* **24**(1): 197-200.

Pilon-Smits, E. (2005). "Phytoremediation." *Annual Review of Plant Biology* **56**(1): 15-39.

Pinyakong, O., H. Habe and T. Omori (2003a). "The unique aromatic catabolic genes in sphingomonads degrading polycyclic aromatic hydrocarbons (PAHs)." *The Journal of General and Applied Microbiology* **49**(1): 1-19.

Pinyakong, O., H. Habe, N. Supaka, P. Pinpanichkarn, K. Juntongjin, T. Yoshida, K. Furihata, H. Nojiri, H. Yamane and T. Omori (2000). "Identification of novel metabolites in the degradation of phenanthrene by *Sphingomonas* sp. strain P2." *FEMS Microbiology Letters* **191**(1): 115-121.

Pinyakong, O., H. Habe, T. Yoshida, H. Nojiri and T. Omori (2003b). "Identification of three novel salicylate 1-hydroxylases involved in the phenanthrene degradation of *Sphingobium* sp. strain P2." *Biochemical and Biophysical Research Communications* **301**: 350 - 357.

Pireddu, L., D. Szafron, P. Lu and R. Greiner (2006). "The Path-A metabolic pathway prediction web server." *Nucleic Acids Research* **34**(suppl_2): W714-719.

Pommerenke, C., I. Gabriel, B. Bunk, R. Münch, I. Haddad, P. Tielen, I. Wagner-Döbler and D. Jahn (2008). "ROSY--a flexible and universal database and bioinformatics tool platform for *Roseobacter* related species." *In Silico Biology* **8**(2): 177-86.

Poretsky, R., I. Hewson, S. Sun, A. Allen, J. Zehr and M. Moran (2009). "Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre." *Environmental Microbiology* **11**(6): 1358-1375.

Portugaly, E., N. Linial and M. Linial (2007). "EVEREST: a collection of evolutionary conserved protein domains." *Nucleic Acids Research* **35**(suppl_1): D241-246.

Pozhitkov, A., P. A. Noble, T. Domazet-Loso, A. W. Nolte, R. Sonnenberg, P. Staehler, M. Beier and D. Tautz (2006). "Tests of rRNA hybridization to microarrays suggest that hybridization characteristics of oligonucleotide probes for species discrimination cannot be predicted." *Nucleic Acids Research* **34**(9): e66.

Pozhitkov, A. E., D. Tautz and P. A. Noble (2007). "Oligonucleotide microarrays: widely applied poorly understood." *Briefings in Functional Genomics* **6**(2): 141-148.

Prak, D. and P. Pritchard (2002). "Degradation of polycyclic aromatic hydrocarbons dissolved in Tween 80 surfactant solutions by *Sphingomonas paucimobilis* EPA 505." *Canadian Journal of Microbiology* **48**(2): 151-158.

Rastogi, G., S. Osman, P. Vaishampayan, G. Andersen, L. Stetler and R. Sani (2010). "Microbial Diversity in Uranium Mining-Impacted Soils as Revealed by High-Density 16S Microarray and Clone Library." *Microbial Ecology* **59**(1): 94-108.

Rehmann, K., N. Hertkorn and A. A. Kettrup (2001). "Fluoranthene metabolism in *Mycobacterium* sp. strain KR20: identity of pathway intermediates during degradation and growth." *Microbiology* **147**(10): 2783-2794.

Relogio, A., C. Schwager, A. Richter, W. Ansorge and J. Valcarcel (2002). "Optimization of oligonucleotide-based DNA microarrays." *Nucleic Acids Research* **30**(11): e51.

Reymond, N., H. Charles, L. Duret, F. Calevro, G. Beslon and J.-M. Fayard (2004). "ROSO: optimizing oligonucleotide probes for microarrays." *Bioinformatics* **20**(2): 271-273.

Rhee, S.-K., X. Liu, L. Wu, S. C. Chong, X. Wan and J. Zhou (2004). "Detection of Genes Involved in Biodegradation and Biotransformation in Microbial Communities by Using 50-Mer Oligonucleotide Microarrays." *Applied and Environmental Microbiology* **70**(7): 4303-4317.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics* **16**(6): 276-277.

Rich, V. I., K. Konstantinidis and E. F. DeLong (2008). "Design and testing of 'genome-proxy' microarrays to profile marine microbial communities." *Environmental Microbiology* **10**(2): 506-521.

Rimour, S., D. Hill, C. Militon and P. Peyret (2005). "GoArrays: highly dynamic and efficient microarray probe design." *Bioinformatics* **21**(7): 1094-1103.

Roesch, L. F. W., R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. H. Daroub, F. A. O. Camargo, W. G. Farmerie and E. W. Triplett (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." *The ISME Journal* **1**(4): 283-290.

Romine, M. F., J. K. Fredrickson and S. M. W. Li (1999a). "Induction of aromatic catabolic activity in *Sphingomonas aromaticivorans* strain F199." *Journal of Industrial Microbiology and Biotechnology* **23**(4): 303-313.

Romine, M. F., L. C. Stillwell, K.-K. Wong, S. J. Thurston, E. C. Sisk, C. Sensen, T. Gaasterland, J. K. Fredrickson and J. D. Saffer (1999b). "Complete Sequence of a 184-Kilobase Catabolic Plasmid from *Sphingomonas aromaticivorans* F199." *Journal of Bacteriology* **181**(5): 1585-1602.

Ronaghi, M. (2001). "Pyrosequencing Sheds Light on DNA Sequencing." *Genome Research* **11**(1): 3-11.

Rosenberg, E., R. Legmann, A. Kushmaro, R. Taube, E. Adler and E. Z. Ron (1992). "Petroleum bioremediation — a multiphase problem." *Biodegradation* **3**(2): 337-350.

Rothmel, R. K., T. L. Aldrich, J. E. Houghton, W. M. Coco, L. N. Ornston and A. M. Chakrabarty (1990). "Nucleotide sequencing and characterization of *Pseudomonas putida catR*: a positive regulator of the *catBC* operon is a member of the LysR family." *Journal of Bacteriology* **172**(2): 922-931.

Rouillard, J.-M., M. Zuker and E. Gulari (2003). "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." *Nucleic Acids Research* **31**(12): 3057-3062.

Royce, T. E., J. S. Rozowsky and M. B. Gerstein (2007). "Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification." *Nucleic Acids Research* **35**(15): e99.

Rychlik, W., W. J. Spencer and R. E. Rhoads (1990). "Optimization of the annealing temperature for DNA amplification *in vitro*." *Nucleic Acids Research* **18**(21): 6409-6412.

Saito, A., T. Iwabuchi and S. Harayama (1999). "Characterization of genes for enzymes involved in the phenanthrene degradation in *Nocardioides* sp. KP7." *Chemosphere* **38**(6): 1331-1337.

Saito, A., T. Iwabuchi and S. Harayama (2000). "A Novel Phenanthrene Dioxygenase from *Nocardioides* sp. Strain KP7: Expression in *Escherichia coli*." *Journal of Bacteriology* **182**(8): 2134-2141.

Saitou, N. and M. Nei (1987). "The Neighbor-Joining method—a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* **4**: 406-425.

Salamonsen, W., K. Mok, P. Kolatkar and S. Subbiah (1999). "BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways." *Pacific Symposium on Biocomputing*: 392-400.

Saleh-Lakha, S., M. Miller, R. Campbell, K. Schneider, P. Elahimanesh, M. Hart and J. Trevors (2005). "Microbial gene expression in soil: methods, applications and challenges." *Journal of Microbiological Methods* **63**(1): 1-19.

Samanta, S. K., O. V. Singh and R. K. Jain (2002). "Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation." *Trends in Biotechnology* **20**(6): 243-248.

Sambrook, J., E. Fritsch and T. Maniatis (2001). **Molecular cloning: A Laboratory Manual - Third Edition**, Cold Spring Laboratory Harbor Press, U.S.A.

Sanger, F., S. Nicklen and A. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463-5467.

Sanguin, H., A. Herrera, C. Oger-Desfeux, A. Dechesne, P. Simonet, E. Navarro, T. M. Vogel, Y. Moënne-Loccoz, X. Nesme and G. L. Grundmann (2006). "Development and validation of a prototype 16S rRNA-based taxonomic microarray for *Alphaproteobacteria*." *Environmental Microbiology* **8**(2): 289-307.

SantaLucia, J. (1998). "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." *Proceedings of the National Academy of Sciences of the United States of America* **95**(4): 1460-1465.

SantaLucia, J. and D. Hicks (2004). "The Thermodynamics od DNA Structural Motifs." *Annual Review of Biophysics and Biomolecular Structure* **33**(1): 415-440.

Sartor, M., J. Schwanekamp, D. Halbleib, I. Mohamed, S. Karyala, M. Medvedovic and C. Tomlinson (2004). "Microarray results improve significantly as hybridization approaches equilibrium." *Biotechniques* **36**(5): 790-796.

Schell, M. A. and E. F. Poser (1989). "Demonstration, characterization, and mutational analysis of NahR protein binding to *nah* and *sal* promoters." *Journal of Bacteriology* **171**(2): 837-846.

Schena, M., D. Shalon, R. Davis and P. Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* **270**(5235): 467-470.

Schloss, P. D. and J. Handelsman (2005). "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness." *Applied and Environmental Microbiology* **71**: 1501-1506.

Schloss, P. D. and J. Handelsman (2006). "Toward a Census of Bacteria in Soil." *PLoS Computanional Biology* **2**(7): e92.

Schmeisser, C., H. Steele and W. Streit (2007). "Metagenomics, biotechnology with non-culturable microbes." *Applied Microbiology and Biotechnology* **75**(5): 955-962.

Scholten, J. C. M., D. E. Culley, L. Nie, K. J. Munn, L. Chow, F. J. Brockman and W. Zhang (2007). "Development and assessment of whole-genome oligonucleotide microarrays to analyze an anaerobic microbial community and its responses to oxidative stress." *Biochemical and Biophysical Research Communications* **358**(2): 571-577.

Schomburg, I., A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn and D. Schomburg (2004). "BRENDA, the enzyme database: updates and major new developments." *Nucleic Acids Research* **32**(suppl_1): D431-433.

Schretter, C. and M. C. Milinkovitch (2006). "OLIGOFAKTORY: a visual tool for interactive oligonucleotide design." *Bioinformatics* **22**(1): 115-116.

Schuler, L., Y. Jouanneau, S. Ní Chadhain, C. Meyer, M. Pouli, G. Zylstra, P. Hols and S. Agathos (2009). "Characterization of a ring-hydroxylating dioxygenase from phenanthrene-degrading *Sphingomonas* sp. strain LH128 able to oxidize benz[*a*]anthracene." *Applied Microbiology and Biotechnology* **83**(3): 465-475.

Scullion, J. (2006). "Remediating polluted soils." *Naturwissenschaften* **93**(2): 51-65.

Sei, K., M. Inaba, R. Upadhye, D. Inoue and M. Ike (2009). "Development of DNA microarray for the evaluation of environmental functions." *Water Science and Technology* **59**(1): 97-107.

Selkov, E., S. Basmanova, T. Gaasterland, I. Goryanin, Y. Gretchkin, N. Maltsev, V. Nenashev, R. Overbeek, E. Panyushkina, L. Pronevitch, E. Selkov, Jr. and I. Yunus (1996). "The metabolic pathway collection from EMP: the enzymes and metabolic pathways database." *Nucleic Acids Research* **24**(1): 26-28.

Selkov, E., Jr., Y. Grechkin, N. Mikhailova and E. Selkov (1998). "MPW: the Metabolic Pathways Database." *Nucleic Acids Research* **26**(1): 43-45.

Seo, J.-S., Y.-S. Keum, Y. Hu, S.-E. Lee and Q. Li (2007). "Degradation of phenanthrene by *Burkholderia* sp. C3: initial 1,2- and 3,4-dioxygenation and meta- and ortho-cleavage of naphthalene-1,2-diol." *Biodegradation* **18**(1): 123-131.

Seo, J.-S., Y.-S. Keum, Y. Hu, S.-E. Lee and Q. X. Li (2006). "Phenanthrene degradation in *Arthrobacter* sp. P1-1: Initial 1,2-, 3,4- and 9,10-dioxygenation, and meta- and ortho-cleavages of naphthalene-1,2-diol after its formation from naphthalene-1,2-dicarboxylic acid and hydroxyl naphthoic acids." *Chemosphere* **65**: 2388-2394.

Seo, J.-S., Y.-S. Keum and Q. X. Li (2009). "Bacterial Degradation of Aromatic Compounds." *International Journal of Environmental Research and Public Health* **6**(1): 278-309.

Sessitsch, A., E. Hackl, P. Wenzl, A. Kilian, T. Kostic, N. Stralis-Pavese, P. Tankouo Sandjong and L. Bodrossy (2006). "Diagnostic microbial microarrays in soil ecology." *New Phytologist* **171**(4): 719-736.

Sigrist, C. J. A., L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch and N. Hulo (2010). "PROSITE, a protein domain database for functional characterization and annotation." *Nucleic Acids Research* **38**(suppl_1): D161-166.

Sigrist, C. J. A., E. De Castro, P. S. Langendijk-Genevaux, V. Le Saux, A. Bairoch and N. Hulo (2005). "ProRule: a new database containing functional and structural information on PROSITE profiles." *Bioinformatics* **21**(21): 4060-4066.

Silva, Í. S., E. d. C. d. Santos, C. R. d. Menezes, A. F. d. Faria, E. Franciscon, M. Grossman and L. R. Durrant (2009). "Bioremediation of a polyaromatic hydrocarbon contaminated soil by native soil microbiota and bioaugmentation with isolated microbial consortia." *Bioresource Technology* **100**(20): 4669-4675.

Silverstein, K. A. T., E. Shoop, J. E. Johnson, A. Kilian, J. L. Freeman, T. M. Kunau, I. A. Awad, M. Mayer and E. F. Retzel (2001). "The MetaFam Server: a comprehensive protein family resource." *Nucleic Acids Research* **29**(1): 49-51.

Singh, B. K. (2010). "Exploring microbial diversity for biotechnology: the way forward." *Trends in Biotechnology* **28**(3): 111-116.

Singleton, D. R., L. Guzman Ramirez and M. D. Aitken (2009). "Characterization of a Polycyclic Aromatic Hydrocarbon Degradation Gene Cluster in a Phenanthrene-Degrading *Acidovorax* Strain." *Applied and Environmental Microbiology* **75**(9): 2613-2620.

Spooner, R. A., K. Lindsay and F. C. Franklin (1986). "Genetic, functional and sequence analysis of the *xylR* and *xylS* regulatory genes of the TOL plasmid pWW0." *Journal of General Microbiology* **132**(5): 1347-1358.

Srinivasiah, S., J. Bhavsar, K. Thapar, M. Liles, T. Schoenfeld and K. E. Wommack (2008). "Phages across the biosphere: contrasts of viruses in soil and aquatic environments." *Research in Microbiology* **159**(5): 349-357.

Staden, R. (1996). "The Staden sequence analysis package." *Molecular Biotechnology* **5**(3): 233-241.

Stenuit, B., L. Eyers, L. Schuler, S. N. Agathos and I. George (2008). "Emerging high-throughput approaches to analyze bioremediation of sites contaminated with hazardous and/or recalcitrant wastes." *Biotechnology Advances* **26**(6): 561-575.

Stolz, A. (2009). "Molecular characteristics of xenobiotic-degrading sphingomonads." *Applied Microbiology and Biotechnology* **81**(5): 793-811.

Story, S. P., E. L. Kline, T. A. Hughes, M. B. Riley and S. S. Hayasaka (2004). "Degradation of Aromatic Hydrocarbons by *Sphingomonas paucimobilis* Strain EPA505." *Archives of Environmental Contamination and Toxicology* **47**(2): 168-176.

Story, S. P., S. H. Parker, S. S. Hayasaka, M. B. Riley and E. L. Kline (2001). "Convergent and divergent points in catabolic pathways involved in utilization of fluoranthene, naphthalene, anthracene, and phenanthrene by *Sphingomonas paucimobilis* var. EPA505." *Journal of Industrial Microbiology and Biotechnology* **26**(6): 369-382.

Story, S. P., S. H. Parker, J. D. Kline, T. R. Tzeng, J. G. Mueller and E. L. Kline (2000). "Identification of four structural genes and two putative promoters necessary for utilization of naphthalene, phenanthrene, fluoranthene by *Sphingomonas paucimobilis* var. EPA505." *Gene* **260**(1-2): 155-169.

Su, Y.-H. and Y.-G. Zhu (2008). "Uptake of selected PAHs from contaminated soils by rice seedlings (*Oryza sativa*) and influence of rhizosphere on PAH distribution." *Environmental Pollution* **155**(2): 359-365.

Suenaga, H., S. Mizuta and K. Miyazaki (2009). "The molecular basis for adaptive evolution in novel extradiol dioxygenases retrieved from the metagenome." *FEMS Microbiology Ecology* **69**(3): 472-480.

Sugawara, H., O. Ogasawara, K. Okubo, T. Gojobori and Y. Tateno (2008). "DDBJ with new system and face." *Nucleic Acids Research* **36**(suppl_1): D22-24.

Sul, W. J., J. Park, J. F. Quensen, III, J. L. M. Rodrigues, L. Seliger, T. V. Tsoi, G. J. Zylstra and J. M. Tiedje (2009). "DNA-Stable Isotope Probing Integrated with Metagenomics for Retrieval of Biphenyl Dioxygenase Genes from Polychlorinated Biphenyl-Contaminated River Sediment." *Applied and Environmental Microbiology* **75**(17): 5501-5506.

Sutherland, J. B. (1992). "Detoxification of polycyclic aromatic hydrocarbons by fungi." *Journal of Industrial Microbiology and Biotechnology* **9**(1): 53-61.

Talla, E., F. Tekaia, L. Brino and B. Dujon (2003). "A novel design of whole-genome microarray probes for *Saccharomyces cerevisiae* which minimizes cross-hybridization." *BMC Genomics* **4**(1): 38.

Tamura, K., J. Dudley, M. Nei and S. Kumar (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0." *Molecular Biology and Evolution* **28**(4): 1596–1599.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). "A Genomic Perspective on Protein Families." *Science* **278**(5338): 631-637.

Teng, Y., Y. Luo, L. Ping, D. Zou, Z. Li and P. Christie (2009). "Effects of soil amendment with different carbon sources and other factors on the bioremediation of an aged PAH-contaminated soil." *Biodegradation*: DOI 10.1007/s10532-009-9291-x.

Terrat, S., E. Peyretaillade, O. Gonçalves, E. Dugat-Bony, F. Gravelat, A. Mone, C. Petit-Biderre, D. Boucher, J. Troquet and P. Peyret (2010). "Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development." *BMC Bioinformatics* **(in press)**.

The UniProt, C. (2010). "The Universal Protein Resource (UniProt) in 2010." *Nucleic Acids Research* **38**(suppl_1): D142-148.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Research* **22**(22): 4673-4680.

Thouand, G., P. Bauda, J. Oudot, G. Kirsch, C. Sutton and J. Vidalie (1999). "Laboratory evaluation of crude oil biodegradation with commercial or natural microbial inocula." *Canadian Journal of Microbiology* **45**(2): 106-115.

van Iersel, M., T. Kelder, A. Pico, K. Hanspers, S. Coort, B. Conklin and C. Evelo (2008). "Presenting and exploring biological pathways with PathVisio." *BMC Bioinformatics* **9**(1): 399.

van Noort, P. C. M., G. Cornelissen, T. E. M. ten Hulscher, B. A. Vrind, H. Rigterink and A. Belfroid (2003). "Slow and very slow desorption of organic compounds from sediment: influence of sorbate planarity." *Water Research* **37**(10): 2317-2322.

Vanbroekhoven, K., A. Ryngaert, L. Bastiaens, P. Wattiau, M. Vancanneyt, J. Swings, R. De Mot and D. Springael (2004). "Streptomycin as a selective agent to facilitate recovery and isolation of introduced and indigenous Sphingomonas from environmental samples." *Environmental Microbiology* **6**(11): 1123-1136.

Vandecasteele, J.-P. (2005). **Microbiologie pétrolière. Concepts, Implications environnementales, Applications industrielles.** Paris, IFP Publications.

Verdick, D., S. Handran and S. Pickett (2002). Key considerations for accurate microarray scanning and image analysis. In G. Kamberova (ed.), DNA array image analysis: nuts and bolts. D. P. LLC, Salem, MA.**:** p. 83-98.

Vernier, P., H. Philippe, P. Samama and J. Mallet (1993). "Bioamine receptors: evolutionary and functional variations of a structural leitmotiv." *EXS* **63**: 297-337.

Viglianti, C., K. Hanna, C. de Brauer and P. Germain (2006). "Removal of polycyclic aromatic hydrocarbons from aged-contaminated soil using cyclodextrins: Experimental study." *Environmental Pollution* **140**(3): 427-435.

Vila, J., J. M. Nieto, J. Mertens, D. Springael and M. Grifoll (2010). "Microbial community structure of a heavy fuel oil-degrading marine consortium: linking microbial dynamics with polycyclic aromatic hydrocarbon utilization." *FEMS Microbiology Ecology* **73**(2): 349-362.

Vlahovicek, K., L. Kajan, V. Agoston and S. Pongor (2005). "The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines." *Nucleic Acids Research* **33**(suppl_1): D223-225.

Wang, X. and B. Seed (2003). "Selection of oligonucleotide probes for protein coding sequences." *Bioinformatics* **19**(7): 796-802.

Waterman, M. (1984). "Efficient sequence alignment algorithms." *Journal of Theoretical Biology* **108**(3): 333-337.

Wernersson, R. and H. B. Nielsen (2005). "OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes." *Nucleic Acids Research* **33**(suppl_2): W611-615.

Wernicke, S. and F. Rasche (2007). "Simple and fast alignment of metabolic pathways by exploiting local diversity." *Bioinformatics* **23**(15): 1978-1985.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko (2008). "Database

resources of the National Center for Biotechnology Information." *Nucleic Acids Research* **36**(suppl_1): D13-21.

Whelan, S., P. I. W. de Bakker and N. Goldman (2003). "Pandit: a database of protein and associated nucleotide domains with inferred trees." *Bioinformatics* **19**(12): 1556-1563.

Wilbur, W. J. (1985). "On the PAM matrix model of protein evolution." *Molecular Biology and Evolution* **2**(5): 434-447.

Williamson, K. E., M. Radosevich and K. E. Wommack (2005). "Abundance and Diversity of Viruses in Six Delaware Soils." *Applied and Environmental Microbiology* **71**(6): 3119-3125.

Williamson, K. E., K. E. Wommack and M. Radosevich (2003). "Sampling Natural Viral Communities from Soil for Culture-Independent Analyses." *Applied and Environmental Microbiology* **69**(11): 6628-6633.

Willison, J. C. (2004). "Isolation and characterization of a novel sphingomonad capable of growth with chrysene as sole carbon and energy source." *FEMS Microbiology Letters* **241**(2): 143-150.

Wu, C. H., L.-S. L. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang and W. C. Barker (2003). "The Protein Information Resource." *Nucleic Acids Research* **31**(1): 345-347.

Wu, L., X. Liu, C. W. Schadt and J. Zhou (2006). "Microarray-Based Analysis of Subnanogram Quantities of Microbial Community DNAs by Using Whole-Community Genome Amplification." *Applied and Environmental Microbiology* **72**(7): 4931-4941.

Wu, L., D. K. Thompson, X. Liu, M. W. Fields, C. E. Bagwell, J. M. Tiedje and J. Zhou (2004). "Development and Evaluation of Microarray-Based Whole-Genome Hybridization for Detection of Microorganisms within the Context of Environmental Applications." *Environmental Science & Technology* **38**(24): 6775-6782.

Xu, D., G. Li, L. Wu, J. Zhou and Y. Xu (2002). "PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis." *Bioinformatics* **18**(11): 1432-1437.

Yakimov, M. M., K. N. Timmis and P. N. Golyshin (2007). "Obligate oil-degrading marine bacteria." *Current Opinion in Biotechnology* **18**(3): 257-266.

Ye, Y. and T. G. Doak (2009). "A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes." *PLoS Computanional Biology* **5**(8): e1000465.

Yergeau, E., S. A. Schoondermark-Stolk, E. L. Brodie, S. Dejean, T. Z. DeSantis, O. Goncalves, Y. M. Piceno, G. L. Andersen and G. A. Kowalchuk (2008). "Environmental microarray analyses of Antarctic soil microbial communities." *The ISME Journal* **3**(3): 340-351.

Ying, L. and M. Sarwal (2009). "In praise of arrays." *Pediatric Nephrology* **24**(9): 1643-1659.

Yona, G., N. Linial and M. Linial (2000). "ProtoMap: automatic classification of protein sequences and hierarchy of protein families." *Nucleic Acids Research* **28**(1): 49-55.

You, I. S., D. Ghosal and I. C. Gunsalus (1988). "Nucleotide sequence of plasmid NAH7 gene *nahR* and DNA binding of the *nahR* product." *Journal of Bacteriology* **170**(12): 5409-5415.

Youssef, N., C. S. Sheik, L. R. Krumholz, F. Z. Najar, B. A. Roe and M. S. Elshahed (2009). "Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys." *Applied and Environmental Microbiology* **75**(16): 5227-5236.

Youssef, N. H. and M. S. Elshahed (2008). "Diversity rankings among bacterial lineages in soil." *The ISME Journal* **3**(3): 305-313.

Yu, C., W. Liu, D. Ferraro, E. Brown, J. Parales, S. Ramaswamy, G. Zylstra, D. Gibson and R. Parales (2007). "Purification, characterization, and crystallization of the components of a biphenyl dioxygenase system from *Sphingobium yanoikuyae* B1." *Journal of Industrial Microbiology and Biotechnology* **34**(4): 311-324.

Yuan, S. and B. Chang (2007). "Anaerobic degradation of five polycyclic aromatic hydrocarbons from river sediment in Taiwan." *Journal of Environmental Science and Health. Part. B, Pesticides, Food Contaminants, and Agricultural Wastes* **42**(1): 63-69.

Zanaroli, G., S. Di Toro, D. Todaro, G. Varese, A. Bertolotto and F. Fava (2010). "Characterization of two diesel fuel degrading microbial consortia enriched from a non acclimated, complex source of microorganisms." *Microbial Cell Factories* **9**(1): 10.

Zanzoni, A., G. Ausiello, A. Via, P. F. Gherardini and M. Helmer-Citterich (2007). "Phospho3D: a database of three-dimensional structures of protein phosphorylation sites." *Nucleic Acids Research* **35**(suppl_1): D229-231.

Zhou, C., M. Lam, J. Smith, A. Zemla, M. Dyer, T. Kuczmarski, E. Vitalis and T. Slezak (2006). "MannDB - A microbial database of automated protein sequence analyses and evidence integration for protein characterization." *BMC Bioinformatics* **7**(1): 459.

Zhou, J., M. A. Bruns and J. M. Tiedje (1996). "DNA recovery from soils of diverse composition." *Applied and Environmental Microbiology* **62**(2): 316-322.

Zhou, N.-Y., J. Al-Dulayymi, M. S. Baird and P. A. Williams (2002). "Salicylate 5-Hydroxylase from *Ralstonia* sp. Strain U2: a Monooxygenase with Close Relationships to and Shared Electron Transport Proteins with Naphthalene Dioxygenase." *Journal of Bacteriology* **184**(6): 1547-1555.

Zhou, N.-Y., S. L. Fuenmayor and P. A. Williams (2001). "*nag* Genes of *Ralstonia* (Formerly *Pseudomonas*) sp. Strain U2 Encoding Enzymes for Gentisate Catabolism." *Journal of Bacteriology* **183**(2): 700-708.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." *Nucleic Acids Research* **31**(13): 3406-3415.

Zylstra, G. J. and E. Kim (1997). "Aromatic hydrocarbon degradation by *Sphingomonas yanoikuyae* B1." *Journal of Industrial Microbiology and Biotechnology* **19**(5): 408-414.

ANNEXES

ANNEXE 1

Protocole de préparation du milieu M457

Milieu M457 :

Composé	Quantité	Poids moléculaire	Fournisseur
Na_2HPO_4	2,44g	358,14 g/mol	Prolabo
KH_2PO_4	1,52g	136,09 g/mol	Prolabo
$(NH_4)_2SO_4$	0,50g	132,14 g/mol	Labosi
$MgSO_4$, $7H_2O$	0,20g	246,47 g/mol	Prolabo
$CaCl_2$, $2H_2O$	0,05g	147,02 g/mol	Prolabo
Solution SL-4*	10,00mL		
H_2O distillée	Qsp 1000mL		

* : voir composition ci-dessous

Afin de solubiliser au mieux les produits, porter à ébullition sous agitation. Après refroidissement, ajuster le pH à 7 puis autoclaver à 120°C pendant 20 minutes.

Solution SL-4 :

Composé	Quantité	Poids moléculaire	Fournisseur
EDTA	0,50g		SIGMA
$FeSO_4$, $7H_2O$	0,20g	278,01 g/mol	Prolabo
Solution SL-6**	100,00mL		
H_2O distillée	Qsp 1000mL		

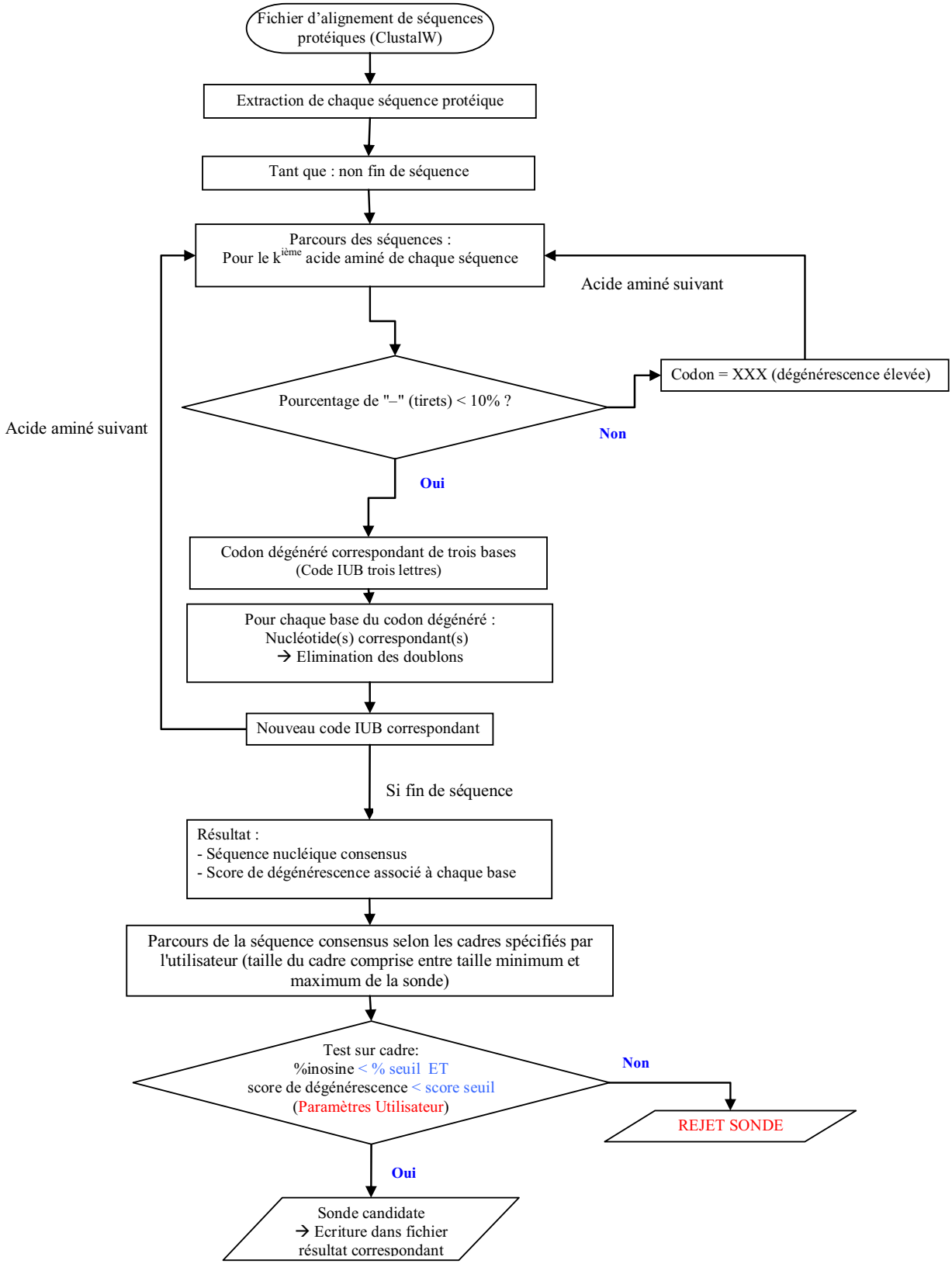
** : voir composition ci-dessous

Solution SL-6 :

Composé	Quantité	Poids moléculaire	Fournisseur
$ZnSO_4$, $7H_2O$	0,10g	287,5 g/mol	Prolabo
$MnCl_2$, $4H_2O$	0,03g	197,9 g/mol	Prolabo
H_3BO_3	0,30g	61,83 g/mol	Prolabo
$CoCl_2$, $6H_2O$	0,20g	237,93 g/mol	Prolabo
$CuCl_2$, $2H_2O$	0,01g	170,48 g/mol	Prolabo
$NiCl_2$, $6H_2O$	0,02g	237,7 g/mol	Riedel-de Häen
Na_2MoO_4, $2H_2O$	0,03g	241,95 g/mol	Prolabo
H_2O distillée	Qsp 1000mL		

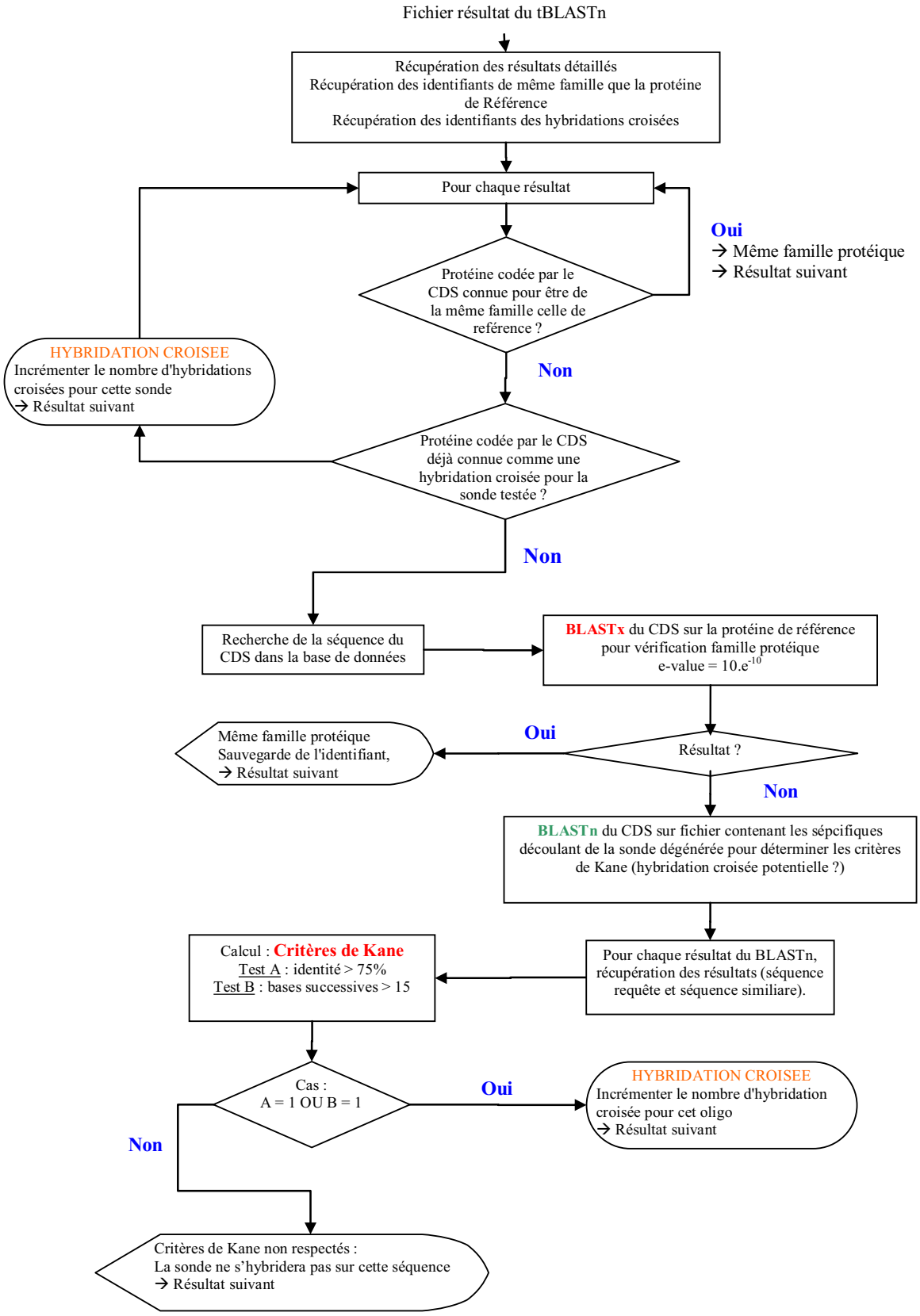
ANNEXE 2

Logigramme d'extraction et de sélection des sondes candidates selon les paramètres de l'utilisateur réalisé par Metabolic Design



ANNEXE 3

Logigramme des étapes pour l'estimation des hybridations croisées potentielles *in silico* des sondes candidates réalisé par Metabolic Design



ANNEXE 4

Paramètre d'E-Value défini pour l'étape de BLASTp de fouille de données pour chaque enzyme de référence et numéros d'accession des enzymes similiares sélectionnées pour l'étape de CLUSTALW et de détermination des sondes

Gène	E-value du BLASTp	Numéros d'accession des enzymes sélectionnées pour réaliser le design des sondes
phnA1a	1e-40	B2Z3Z2, A2TC87, A4XDY3, O85843, A9Y004, B5L7S0, B5L7R9, Q1HCP6, Q7WUA0
phnA2a	1e-30	A2TC88, A4XDY2, O85842, B5L7R8
ahdA1c	1e-40	A2TC29, A9XZZ2, Q65AS5
ahdA2c	1e-30	A9XZZ3, Q65AS6, A4XDV1, O85992, A2TC30, Q9Z4T6
bphB	1e-40	Q14RW3, O85972
bphC	1e-40	P11122, Q6LCU9, Q7DG81, A4XDU9, O85990, A9XZZ5, Q65AS8, Q9KWI2
bphA3	1e-20	O34128, Q65AS7, A9XZZ4, A4XDV0, O85991, Q83VL0
ahdA4	1e-40	Q83VI9, A4XDS3, O85962
nahE	1e-40	O85960, A4XDS1, A2TC61, Q83VI8, Q9X9Q6
ahdA1d	1e-40	A4XDU5, O85986, A9XZZ9, A2TC36, Q7WUA1
ahdA2d	1e-30	A2TC37, A4XDU4, O85985
bphA1a	1e-40	A4XDS5, A2TC57, Q83VJ1
bphA2a	1e-30	A4XDS6
bphA1b	1e-40	Q75WN5, Q934B6, Q6REQ7, B8X9W8, A2TC55, Q83VJ3
bphA2b	1e-30	A2TC54, Q83VJ4
bphA1e	1e-40	A2TC62, Q83VI7
bphA2e	1e-30	A2TC63, Q83VI6
nahD	1e-40	A4XDV3, O85994, Q65AS4, A2TC28
xylX	1e-40	A4XDU8, O85989, A9XZZ6, A2TC33
xylY	1e-30	A4XDU7, O85988, A9XZZ7, A2TC34
bphK	1e-40	O33705, A2TC38, Q83VK4, A4XDU3, O85984, O30347, A9Y001
bphF	1e-40	A4JPX7, A1T2N6, A5V856, A5V6T8, A5VGU4
bphR	1e-40	A4XDS4, A2TC58
xylA	1e-40	A2TC52
xylM	1e-40	Q83VJ5, O30865, P21395
xylC	1e-40	A2TC49, Q402C7, Q9X9Q8, Q9Z3W8, B4EDN3
xylE	1e-40	Q79BB1, Q798L3, Q52444, A9UCR4, Q9KWI1, Q0GC10, Q8RMH9, Q83VK2, Q45459, Q19A88, B2Z3Z1, Q9ZAY3
nahF	1e-40	B8KG99
nagG	1e-40	A1K8H8, B1Y192, Q12EV5, B7YLP6, O87618, Q5GRC1, Q8GHI9, A6VYQ3, A1VQ64, Q3S4D3, Q45693, A3RZ19, B5SMB7, B5S694, Q8Y0F4
nagH	1e-30	Q79BZ1, Q9Z5Q7, Q3S4D2
nagR	1e-30	Q7WT50, Q8VUD7, Q7WT51, Q7WT52, A1VQ66, Q3S4D5
catA	1e-40	A4XFH8
catR	Aucune sélectionnée	
phdI	1e-40	O24721, Q9FBF3, B0BK98, B0BK99
Orf007	1e-20	A4XDX1
Orf158	1e-20	A3WDH0
Orf569	1e-20	A4XE42
Orf597	1e-20	A4XE48
Orf758	1e-20	A4XE71
gyrB	1e-80	A3XDH8, A4ETT8, A3S8M5, A3STI7, A3K334, A3V0Y7, A3SKA1, A3W428, A3JM17, A3PFL9, A4WNF4, A3VCU2, A1AZ36, A0NUF7, A3WSP3, A1UUE0, A3UIJ1, A3VNH4, A3WH32, A1JT76, A4TGL8, A1AHN4, A1S1H2, A0L3J2

ANNEXE 5

Caractéristiques des sondes dégénérées définies avec Metabolic Design pour chacun des 40 gènes ciblés

Ce tableau présente le nom du gène ciblé, le nom de la sonde dégénérée, la séquence dégénérée de chacune des sondes sélectionnées, le nombre de sondes spécifiques découlant de chaque sonde dégénérée, et enfin la position des sondes sur la séquence du gène de référence codant l'enzyme utilisée pour la fouille de données via l'étape du BLASTp. Nomenclature : **M** : A et C ; **K** : G et T ; **R** : A et G ; **W** : A et T ; **S** : G et C ; **Y** : C et T ; **V** : A, C et G ; **H** : A, C et T ; **D** : A, G et T ; **B** : G, T et C ; **I** : A, C, G et T.

Gène	Nom de la sonde	Séquence de la sonde	Nombre de sondes spécifiques	Positions sur le gène de référence
phnA1a	phnA1a_MD_A	GTITGYAAYTAYCAYGGITGGGT	256	294 – 316
	phnA1a_MD_B	CAYGARATHGARGTITGGACITA	384	957 – 979
phnA2a	phnA2a_MD_A	GARGAYATHCAYTAYTGGATGCC	48	123 – 145
	phnA2a_MD_B	GGICARGTITGGATGGARGAYCC	128	261 – 284
ahdA1c	ahdA1c_MD_A	GARTGYGTITAYCAYCARTGGGC	128	318 – 340
	ahdA1c_MD_B	GAYGCIGCIGAYAARCARGCITA	1024	771 – 793
ahdA2c	ahdA2c_MD_A	GAYGAYMGIYTIGARGARTGGCC	1024	081 – 103
	ahdA2c_MD_B	ATHGAYACIATGATGGTIMGICC	768	459 – 481
bphB	bphB_MD_A	AAYGTIGGIATHTGGGAYTWYAT	768	261 – 283
	bphB_MD_B	AAYBTIAARGGITAYTTYTTYGG	384	348 – 370
bphC	bphC_MD_A	CCITAYTTYATGCAYTGYAAYGA	128	558 – 580
	bphC_MD_B	TGGYTITGGGARTTYGGITGGGG	128	777 – 799
bphA3	bphA3_MD_A	ATHATHGARTGYCCITTYCAYGG	576	180 – 202
	bphA3_MD_B	ATHGAIGAYGGITGGGTITGYAT	768	279 – 302
ahdA4	ahdA4_MD_A	GCIAAYGTICCIGAYAAYTTYTT	1024	159 – 181
	ahdA4_MD_B	CARGARACITAYCARAAYGCIGC	512	867 – 889
nahE	nahE_MD_A	GARGCITTYAARTTYGAYTTYCC	256	459 – 481
	nahE_MD_B	CCIATHGAYTTYGAYTAYTAYGG	384	597 – 619
ahdA1d	ahdA1d_MD_A	TGYGTITAYCAYCARTGGGCITA	256	309 – 331
	ahdA1d_MD_B	ATGGARGAYGGIGARGCIGTIGA	512	1068 – 1090
ahdA2d	ahdA2d_MD_A	CCIGCIGCIGTIATGTAYTGYGA	1024	153 – 175
	ahdA2d_MD_B	GCIAAYGTITTYCCIGARCAYTT	1024	219 – 241
bphA1a	bphA1a_MD_A	TTYACITGYAAYTAYCAYGGITG	512	294 – 316
	bphA1a_MD_B	GGICARATHGARCARTGGACITG	384	972 – 994
	bphA1a_MD_C	CARATGGAYGTITAYACIAAYAT	256	1146 – 1168
bphA2a	bphA2a_MD_A	GAYAAYGARGAYTTYGARGGITG	256	84 – 106
	bphA2a_MD_B	TGGGCIGARGAYCCICCIAAYTA	512	270 – 292
bphA1b	bphA1b_MD_A	TTYGTITGYAAYTAYCAYGGITG	512	315 – 337
	bphA1b_MD_B	TGGAARTTYGGIGTIGARAAYTT	256	606 – 628
	bphA1b_MD_C	GARATGGAYGAYGGIGARAAYTG	128	1077 – 1099
bphA2b	bphA2b_MD_A	CAYTAYCAYATGCCIGGIATHGA	384	153 – 175
	bphA2b_MD_B	ACIMGIATGGCITAYTAYAAYGA	1024	219 – 241
bphA1e	bphA1e_MD_A	ACIGARTTYGARTGYCCITAYCA	512	288 – 310
	bphA1e_MD_B	GARGTICAYTAYGCITAYTTYGC	512	882 – 904
bphA2e	bphA2e_MD_A	TGGGYIGGIACITTYCARGAYTA	1024	234 – 256
nahD	nahD_MD_A	TGGCCIATHGAYATHCCIGARGC	576	108 – 130
	nahD_MD_B	ATHTGGGGICARGGIATHGAYCC	576	342 – 364
xylX	xylX_MD_A	CAYGCIAAYTAYTTYATGACIGT	256	702 – 724
	xylX_MD_B	TAYGTIGCIATHCAYGAYGARTG	768	1311 – 1333
xylY	xylY_MD_A	GCIYTIGAYGAYAARGAYTGGGA	512	60 – 82
	xylY_MD_B	GARTAYTGGGYICCIGCITGGGA	512	114 – 136
bphK	bphK_MD_A	YTITAYATHGCIGAYCARAAICC	768	210 – 232
	bphK_MD_B	GARTTYCAYAARGCITTYGTICC	512	309 – 331
bphF	bphF_MD_A	ATGGTIGGIGGICARGARGAYAT	512	951 – 973
bphR	bphR_MD_A	AAYTGYGCIGCIATHCCIGARGG	768	801 – 823

ANNEXE 5 (suite)

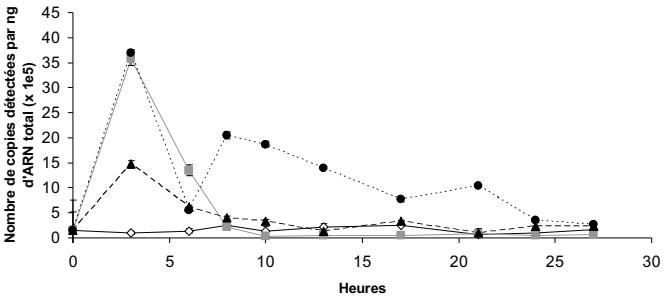
Gène	Nom de la sonde	Séquence de la sonde	Nombre de sondes spécifiques	Positions sur le gène de référence
xylA	xylA_MD_A	YTIGCIATGCCICAYGAYTGYAA	1024	114 – 136
	xylA_MD_B	ATGCCICAYGAYTGYAARGTIGG	256	120 – 140
xylM	xylM_MD_A	GTIGARGGITTYAAYTAYTTYCA	512	732 – 754
	xylM_MD_B	CAYCAYGCITGGAAYCAYHTIGG	768	798 – 820
xylC	xylC_MD_A	CARATHATHCCITGGAAYGTICC	576	471 – 493
xylE	xylE_MD_A	GCITTYTTYCARGCITTYGAYGA	256	123 – 145
	xylE_MD_B	AAYAARGCICAYGAYGTIGCITT	1024	585 – 607
nahF	nahF_MD_A	GARGCIGGIATGGTICAYATHAA	768	1269 – 1291
nagG	nagG_MD_A	GGIAAYTGGAARYTIATGCARGA	256	627 – 649
	nagG_MD_B	TTYGAYTTYGTITGGACICAYTT	256	996 – 1018
nagH	nagH_MD_A	TGGGARAARTGGCCIGAITTYTT	128	075 – 097
nagR	nagR_MD_A	ACIGAYATHGGIGARATGTAYTT	384	309 – 331
catA	catA_MD_A	CAYGTICAYTTYTTYGTIGARGC	512	660 – 682
	catA_MD_B	GCIGAYGAYTTYGCITTYGGIAC	1024	738 – 760
catR	catR_MD_A	CCIAARACIATGCARCARTAYGC	256	234 – 256
	catR_MD_B	GAYATHTGYTAYGTICCIATHGC	1152	759 – 781
phdI	phdI_MD_A	GGICARTGGATHGCIGARGARCA	384	90 – 112
	phdI_MD_B	ATGSCIACIATGGAYTGYTGGGT	128	789 – 811
Orf007	Orf007_MD_A	ATGGARCAYCARGCIACIGARMG	512	045 – 067
	Orf007_MD_B	GTIGARCARAAYGTICARAAYGG	512	264 – 286
Orf158	Orf158_MD_A	GCIAAYGAYATHGCIACITTYAA	384	390 – 412
Orf569	Orf569_MD_A	ATHGAYGAYGCIGTICARGARAC	768	147 – 169
Orf597	Orf597_MD_A	GAYTGYGGIGCIAARATGAARAT	256	213 – 235
	Orf597_MD_B	GAYTAYCCIATGGTIGCICCIGA	1024	309 – 331
Orf758	Orf758_MD_A	TTYCAYGTIGARATHATGMGIAT	768	402 – 424
	Orf758_MD_B	GCICCIACITTYGARTTYGARGG	1024	693 – 715
gyrB	gyrB_MD_A	TAYATHGGIGAYACIGAYGAYGG	768	102 – 124

ANNEXE 6

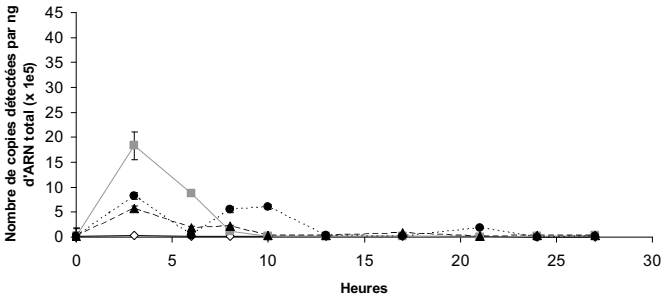
Suivis d'expression des gènes d'intérêt de la souche modèle EPA505 par une approche de PCR quantitative en présence de différentes sources carbonées

Pour chacun des gènes étudiés (*phnA1a, phnA2a, ahdA1c, ahdA2c, bphB, bphC, bphA3* et *ahdA4* ; respectivement de A à H) sont représentés les profils d'expression obtenus pour les différentes cinétiques analysées. Légende : losange blanc : glucose, carré gris : phénanthrène, triangle noir : fluoranthène, rond noir : phénanthrène et fluoranthène.

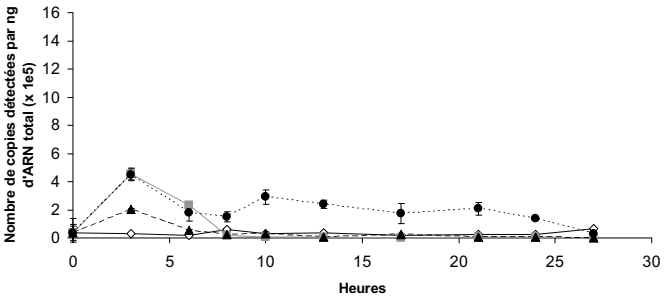
(A) *phnA1a*



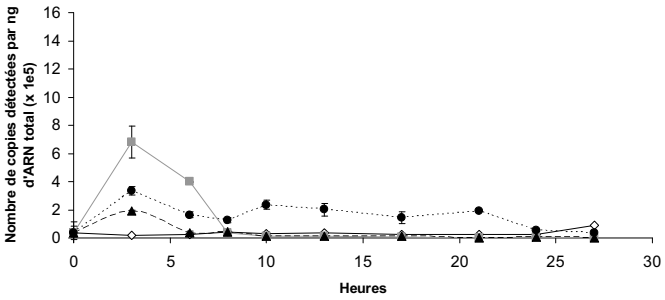
(B) *phnA2a*



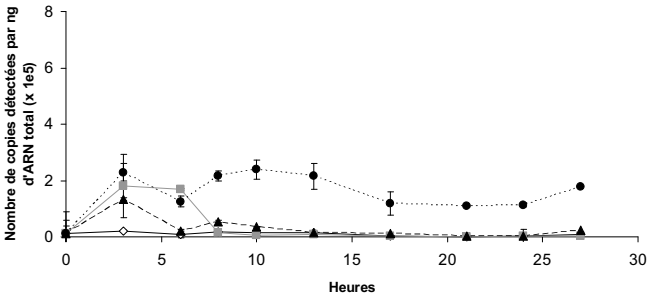
(C) *ahdA1c*



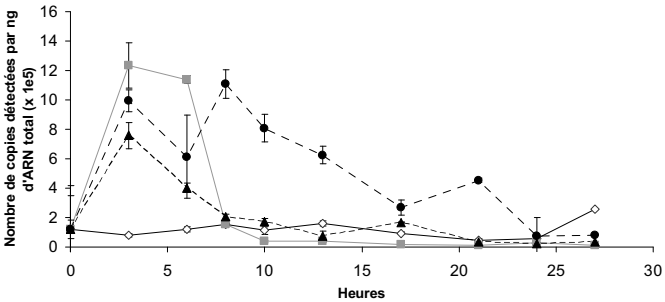
(D) *ahdA2c*



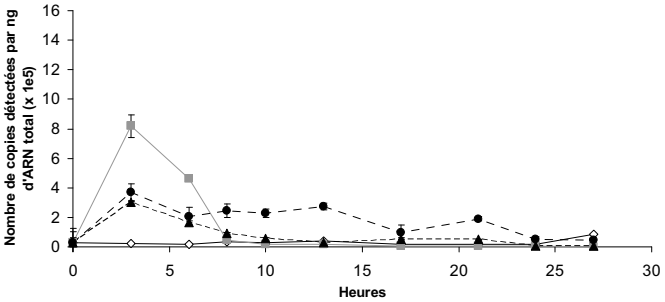
(E) *bphB*



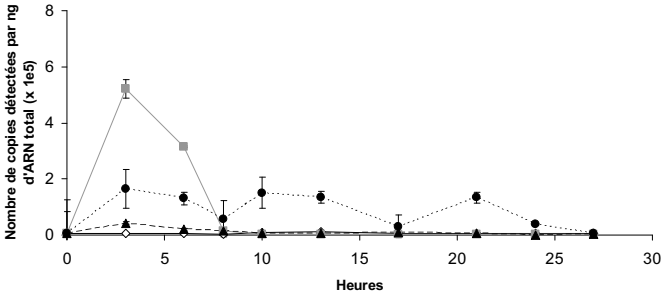
(F) *bphC*



(G) *bphA3*



(H) *ahdA4*



ANNEXE 7

Publication acceptée dans la revue BMC Bionformatics

Terrat et al. BMC Bioinformatics 2010, 11:478
http://www.biomedcentral.com/1471-2105/11/478

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development

Sébastien Terrat[1,2,3], Eric Peyretaillade[1,2], Olivier Gonçalves[1,2], Eric Dugat-Bony[2,3], Fabrice Gravelat[2,3], Anne Moné[2,3], Corinne Biderre-Petit[2,3], Delphine Boucher[1,2], Julien Troquet[4], Pierre Peyret[1,2*]

Abstract

Background: Microorganisms display vast diversity, and each one has its own set of genes, cell components and metabolic reactions. To assess their huge unexploited metabolic potential in different ecosystems, we need high throughput tools, such as functional microarrays, that allow the simultaneous analysis of thousands of genes. However, most classical functional microarrays use specific probes that monitor only known sequences, and so fail to cover the full microbial gene diversity present in complex environments. We have thus developed an algorithm, implemented in the user-friendly program Metabolic Design, to design efficient explorative probes.

Results: First we have validated our approach by studying eight enzymes involved in the degradation of polycyclic aromatic hydrocarbons from the model strain *Sphingomonas paucimobilis* sp. EPA505 using a designed microarray of 8,048 probes. As expected, microarray assays identified the targeted set of genes induced during biodegradation kinetics experiments with various pollutants. We have then confirmed the identity of these new genes by sequencing, and corroborated the quantitative discrimination of our microarray by quantitative real-time PCR. Finally, we have assessed metabolic capacities of microbial communities in soil contaminated with aromatic hydrocarbons. Results show that our probe design (sensitivity and explorative quality) can be used to study a complex environment efficiently.

Conclusions: We successfully use our microarray to detect gene expression encoding enzymes involved in polycyclic aromatic hydrocarbon degradation for the model strain. In addition, DNA microarray experiments performed on soil polluted by organic pollutants without prior sequence assumptions demonstrate high specificity and sensitivity for gene detection. Metabolic Design is thus a powerful, efficient tool that can be used to design explorative probes and monitor metabolic pathways in complex environments, and it may also be used to study any group of genes. The Metabolic Design software is freely available from the authors and can be downloaded and modified under general public license.

Background

Assessing the metabolic potential of microorganisms in variable ecosystems is a novel and stimulating challenge in biology. Microorganisms are present in all environmental habitats, even the most extreme, yet despite their ubiquity, we know relatively little about these communities. Microorganisms display vast diversity, each one having its own set of genes, cell components and metabolic reactions [1]. Thus 1 g of soil may contain up to 10^9 bacteria cells, which may represent between 1,000 and 10,000 different species [2,3]. Assuming 3,000 genes per single bacteria genome, there will thus be up to 3×10^{12} genes mediating huge and various biological processes [3,4]. To overcome the limits of cultivation, several high throughput approaches have been developed to explore genetic contents, such as metagenomics or DNA microarrays [1,5,6]. Numerous random shotgun metagenomic projects have caused the publicly available sequence data to increase exponentially, giving us a basis to study complex ecosystems [1,5]. In some cases,

* Correspondence: pipeyret@univ-bpclermont.fr
[1]Clermont Université, Université d'Auvergne, Laboratoire: Microorganismes Génome et Environnement, BP 10448, F-63000 CLERMONT-FERRAND, France
Full list of author information is available at the end of the article

these sequence data were used to identify different species in environmental or clinical samples with DNA microarrays [5]. Moreover, these data should improve our knowledge not only of genome organization and genome evolution but also of biological processes and biological activities. However, although such sequencing approaches can rapidly generate large amounts of data, they give only a snapshot of genetic information and can be laborious and costly when complex ecosystems are to be studied. Also, DNA sequencing is not informative on gene expression and regulation. Metatranscriptomic studies are promising, but several obstacles have to be crossed before they can be widely used [1,7]. Indeed, sequencing approaches highlight the difficulties of accurate functional annotation of unknown proteins without experimental data; unsupervised annotation of proteins by software pipelines suffers from very high error rates. Spurious functional assignments are usually caused by species homology-based transfer of information from existing database entries to new target sequences [8,9]. Such functional annotation errors are due to local similarities between the query and functionally annotated sequences. Hence, two protein sequences may have two different biological functions, but a same protein domain. This approach, based on homologous gene prediction, presents another major drawback: it can fail to identify novel enzymes that have the same function, but a different primary structure from known enzymes [10]. Today, the main sources for such protein sequence data are Swiss-Prot, TrEMBL and GenPept. This last should be considered as an equivalent to the Swiss-Prot/TrEMBL databases with a high level of redundancy in terms of protein sequences [11]. Unlike TrEMBL, the Swiss-Prot database contains curated datasets of high quality [12].

Another high throughput tool, functional DNA microarrays, can also be used for monitoring metabolic diversity of microbial populations in environmental samples. In a single experiment, thousands of genes can be simultaneously detected. Several studies already demonstrate the usefulness of functional DNA microarrays for exploring various ecosystems [13-15]. Hybridization of microarrays with mRNA targets permits low-cost, easy quantitative estimates of gene expression levels [16]. Monitoring environmental metabolic processes can be made more powerful, and so more useful, by designing explorative probes to ensure the detection of genes not already discovered and deposited in databases. However, microarray probe design software determines specific probes to monitor only known sequences [17]. Thus only a small fraction of genes encoding microbial enzymes can be studied with these probes. To solve this problem, degenerate probes need to be defined, as for PCR-based applications [18].

Probe design also has to allow for the constraints of cross-hybridization. Specificity is a measure of the inability of a probe to bind strongly to non-target sequences that may be present in a biological sample. This can be accomplished by avoiding probes with excessive sequence similarity to a non-target sequence that may be present during the hybridization [19,20]. These problems of cross-hybridization emphasize the need to take into account the fact that the studies are conducted on complex environments. As thermodynamic constraints are not yet completely understood [21], sequence similarity is currently the prime parameter used to check probe specificity. A previously reported and extensively cited work by Kane and coworkers [22] on 50-mer probes, shows that a probe must meet two conditions to be specific: (*i*) the oligonucleotide sequence must have no more than 75% similarity (among all sequences) with a non-targeted sequence present in the hybridization pool, and (*ii*) the oligonucleotide sequence must not include a stretch of identical sequence longer than 15 contiguous bases.

Here we describe a new algorithm, implemented in a user-friendly program, named Metabolic Design, which will generate efficient explorative probes using a simple convenient graphical interface. The practical utility of this approach was demonstrated by studying several genes encoding enzymes involved in the degradation of diverse polycyclic aromatic hydrocarbons from the model strain *Sphingomonas paucimobilis* sp. EPA505 (strain EPA505) and assessing metabolic capacities of microbial communities in a soil contaminated with aromatic hydrocarbons.

Results

The Metabolic Design software

Our aim is to build a graphical display of given biological processes and perform exhaustive sequence mining of all available protein sequences for each biological step studied. The graphical user interface (GUI) allows for example the graphical reconstruction of tailor-made metabolic pathways, with metabolites and enzymes represented respectively with nodes and edges (Figure 1). Using appropriate keywords, correctly annotated protein sequences are extracted from a curated database (by default Swiss-Prot potentially enriched with personal data) for each edge of the graph. The user can freely select the most suitable protein as a reference sequence query. This sequence is then used to carry out exhaustive mining of similar proteins from public and/or personal databases. The strategy of probe design using Metabolic Design software is described in Figure 2 and detailed in Methods under 'Software implementation'.

In our study, reference sequences are extracted from the highly curated database Swiss-Prot formatted for the

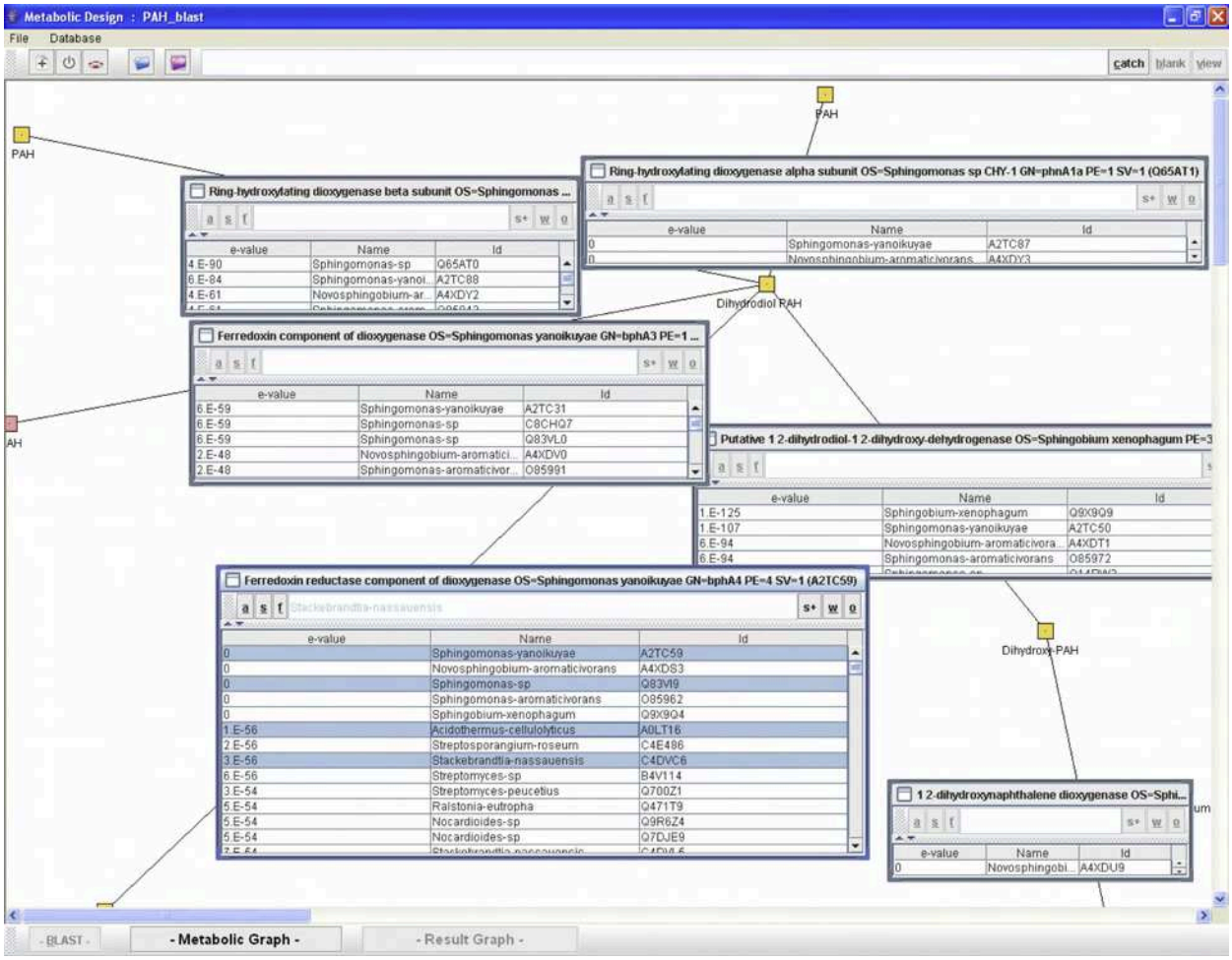


Figure 1 Results window produced by Metabolic Design. Each metabolite is represented by small yellow squares, called nodes, and enzymes as edges between nodes. Inner windows give the parsing of BLASTp results ordered by increasing Expected Value and obtained for each reference protein as query. For each extracted homologous protein sequence, data such as sequence in EMBL format (f button), sequence in FASTA format (s button), or split BLASTp alignment results (a button) are directly available through the inner window toolbar buttons. The w button, allows the execution of ClustalW alignment on pre-selected protein sequences. Such sequences can also be saved in a single file in FASTA format (s+ button), and/or used to launch the probe design module (o button). Additional functions have also been implemented. The user can automatically highlight potential metabolic capacities of a given organism (species name) subsequently using the 'catch' and 'view' buttons at the top of the window.

application to ensure efficient mining. In addition, the reference database is enriched with TrEMBL protein sequences biologically validated when a non-studied orthologous protein sequence was found in the Swiss-Prot database. For every listed protein, data are extracted using a homologous approach with a BLASTp program against concatenated Swiss-Prot and TrEMBL databases. Thus the selection of candidate protein sequences based on similarity criteria bypasses functional annotation errors. Extracted sequences are then automatically filtered and displayed in graphical edges for each studied enzyme. The results are finally organized according to increasing expected value, or organism origin, and miscellaneous functions are also implemented in the toolbar to facilitate additional data extraction and visualization (Figure 1).

This multiple alignment is then used to design specific explorative oligonucleotide probes targeting studied proteins (Figure 3A), using the following procedure. To reduce insertion-deletion (indel) regions in multiple alignments, a first filtering step is carried out to exclude sequences with high size divergence compared with the reference query. A degenerate nucleic consensus sequence based on the IUPAC (International Union of Pure and Applied Chemistry) nomenclature is defined from the protein multiple alignment using the backtranslation approach [23]. For each molecular site, potential amino acids are backtranslated taking into account all

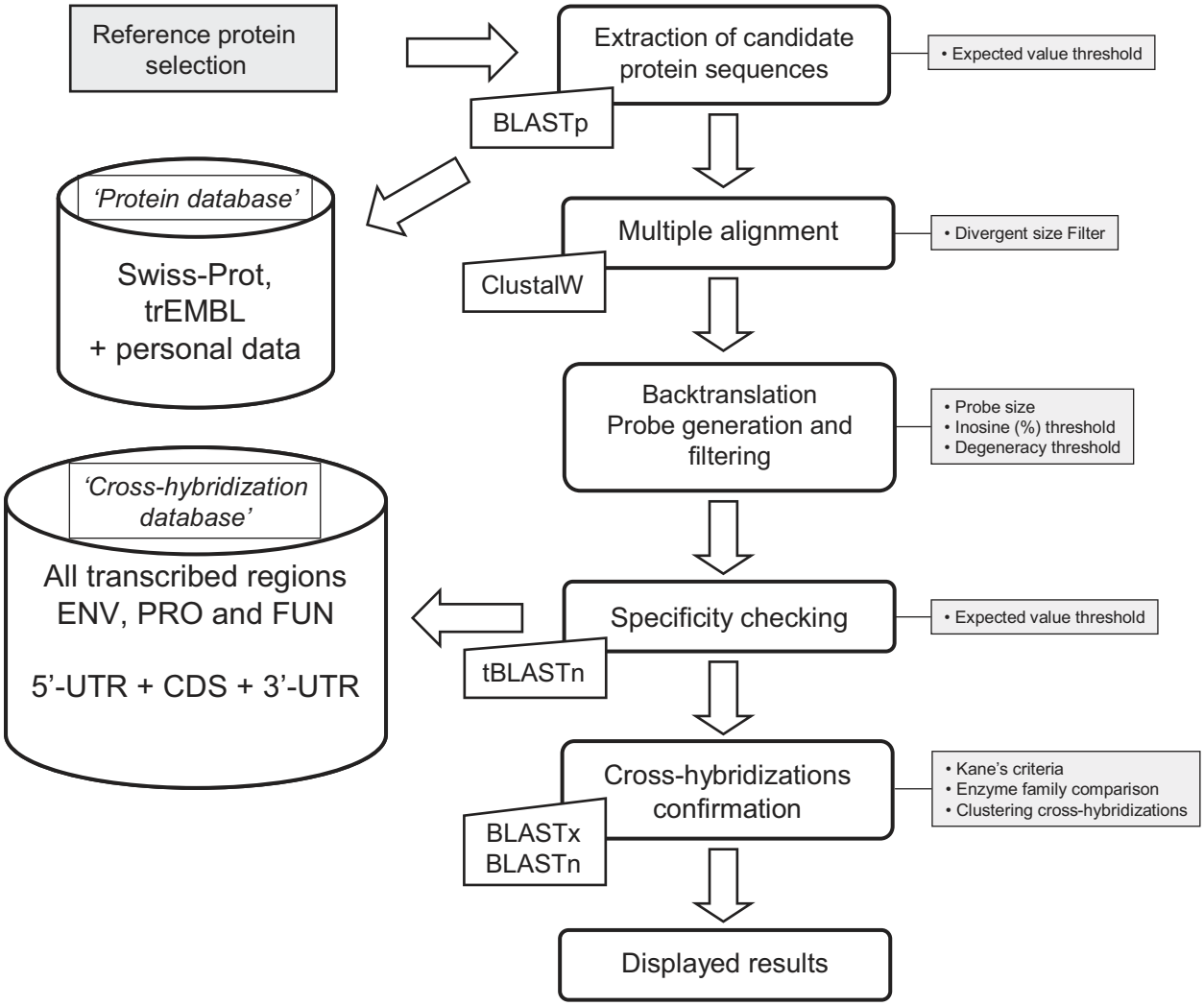


Figure 2 Strategy to design explorative probes for functional microarrays used in Metabolic Design. After extraction of potential candidate sequences by BLASTp using query reference protein to compare against concatenated Swiss-Prot and TrEMBL databases, a multiple alignment with selected protein sequences and the reference protein is performed. The next step is in two parts: (*i*) for each molecular site, amino acids are backtranslated, taking into account all genetic code redundancy to determine a degenerate nucleic consensus sequence, (*ii*) probes are then extracted from this consensus sequence, according to defined user parameters. The program then searches for all potential cross-hybridizations for each selected probe against the 'Cross-hybridization database' by tBLASTn. Kane's criteria are then checked for all positive results by BLASTn. If Kane's criteria are in agreement with a potential cross-hybridization, the program also checks whether it is a potential member of the targeted enzyme family using a BLASTx comparison against the reference protein. Cross-hybridization results are then clustered by BLASTn, stored and visualized in an output file.

genetic code redundancy (Figure 3B). Probes are then extracted from this consensus sequence, according to three defined user parameters: probe size, degeneracy and inosine composition thresholds. Along the consensus sequence, the algorithm extracts all probes by incrementing the constant defined probe size in a window. All probes with degeneracy and inosine composition under the set thresholds are then listed in an inner window in the GUI. Thus the user can select all or some pre-selected potential probes for specificity testing. To reduce computing time, during this test the algorithm generates all peptide combinations for each degenerate probe. Thus the degeneracy code redundancy is bypassed and the number of comparisons is greatly reduced. This test is carried out using tBLASTn against the 'Cross-hybridization database' using Kane's algorithm criteria [22]. Indeed, those parameters are used to check all positive results by comparison at the nucleotide level with BLASTn. If those criteria are in agreement with a potential cross-hybridization this may also reflect hybridization with a member of the targeted enzyme family. To avoid this bias, the algorithm extracts

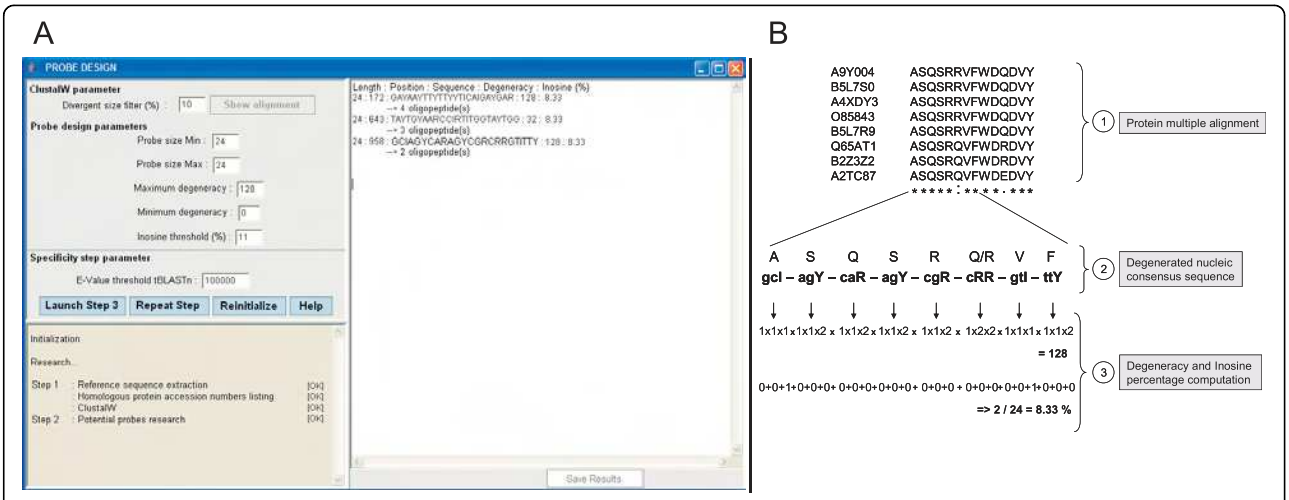


Figure 3 Probe determination with Metabolic Design. A: The inner window of Metabolic Design showing results for designing probes with Metabolic Design and parameters (such as probe size, degeneracy or inosine composition) defined by the user. The user's chosen parameters are visible on the left, and potential probes are listed on the right. The program also displays all potential peptide combinations for each degenerate probe (named as oligopeptides) with probe listing. B: Example of probe design approach, degeneracy calculation and inosine percentage determination for the third probe in the inner window. Note that inosine residues are not taken into account for the degeneracy calculation step.

the complete sequence of the gene harboring the potential cross-hybridization region and compares it with the reference protein using the BLASTx program. Finally, a file containing all potential cross-hybridizations for every candidate probe is automatically clustered, created and displayed.

Data mining and probe selection for microarray experiments using Metabolic Design

To validate our probe design strategy, we focus on metabolic pathways involved in the biodegradation of polycyclic aromatic hydrocarbons (PAHs). PAHs are a class of fused-ring aromatic compounds that are ubiquitous environmental pollutants known to be toxic, mutagenic and/or carcinogenic. Many researchers have therefore focused on the biodegradation of these pollutants by microorganisms, especially bacteria. Several enzymes involved in these critical biodegradation steps have been characterized and their sequences deposited in databases [24-28].

In this study, we choose to target eight genes (*phnA1a*, *phnA2a*, *bphC*, *bphA3*, *ahdA1c*, *ahdA2c*, *ahdA4* and *bphB*) (Table 1) known to be involved in the degradation of several PAHs (such as phenanthrene (PHE) and fluoranthene (FLA)). Using our defined data mining strategy, we first construct the metabolic pathway with respective substrates and products of each metabolic step. Secondly, for each of these metabolic steps, one reference enzyme is extracted from our curated database (Swiss-Prot and trEMBL validated data). Homologous proteins are then retrieved from complete databases (Swiss-Prot and TrEMBL). Based on defined expected threshold values, different sequences are selected (Table 1) and multiple alignments are then performed to ensure probe design step.

To improve our probe design, we have applied two different strategies, using the same multiple alignments. In these strategies, we set the probe length at 24-mer, representing the best compromise between probe specificity and sensitivity criteria [29]. In the first strategy, (degeneracy threshold: 129, inosine threshold: 25%), we have determined a first set of probes for each targeted enzyme. However, owing to high percentages of inosine, these probes generally show a high degree of total degeneracy. Indeed, like the inosine residues are not taken into account for the degeneracy threshold determination, probes may present a maximum total degeneracy of 528,384. To reduce the number of potential specific probes derived from each degenerate defined probe, a second strategy with more stringent parameters is applied (degeneracy threshold: 258, inosine threshold: 9%, maximum total degeneracy of 4 128). Using these parameters, we have found another set of probes for each targeted enzyme. We then choose among the two probe sets obtained with these two strategies, the best probes based on several sequentially evaluated criteria: (1) the total number of potential cross-hybridizations to decrease possibilities of non-homologous hybridizations, (2) the probe total degeneracy (including inosine composition) to restrict the number of specific probes in the microarray, and (3) the position of each probe in the reference sequence to target different regions for each

Table 1 Reference enzyme information

		REFERENCE PROTEIN			BLAST THRESHOLD AND SEQUENCES USED	
Gene	**Enzyme**	**Organism**	**Accession Number**	**Reference**	**BLASTp e-value**	**Chosen enzymes for the probe design**
phnA1a	Putative alpha subunit of ring-hydroxylating dioxygenase	*Sphingomonas* sp. CHY-1	Q65AT1	[24]	1e-40	B2Z3Z2, A2TC87, A4XDY3, 085843, A9Y004, B5L7S0, B5L7R9, Q1HCP6, Q7WUA0
phnA2a	Putative beta subunit of ring-hydroxylating dioxygenase	*Sphingomonas* sp. CHY-1	Q65AT0	[24]	1e-30	A2TC88, A4XDY2, 085842, B5L7R8
ahdA1c	Putative large subunit of oxygenase	*Sphingomonas* sp. P2	Q83VL2	[27,32]	1e-40	A2TC29, A9XZZ2, Q65AS5
ahdA2c	Putative small subunit of oxygenase	*Sphingomonas* sp. P2	Q83VL1	[27,32]	1e-30	A9XZZ3, Q65AS6, A4XDV1, 085992, A2TC30, Q9Z4T6
bphB	Putative 1, 2-dihydrodiol-l, 2-dihydroxy-dehydrogenase	*Sphingobium xenophagum*	Q9X9Q9	[25]	1e-40	Q14RW3, 085972
bphC	Putative biphenyl-2,3-diol 1,2-dioxygenase	*Sphingobium xenophagum*	P74836	[25]	1e-40	PI 1122, Q6LCU9, Q7DG81, A4XDU9, 085990, A9XZZ5, Q65AS8, Q9KWI2
bphA3	Putative ferredoxin component of dioxygenase	*Sphingomonas yanoikuyae*	A2TC31	[24,32]	1e-20	034128, Q65AS7, A9XZZ4, A4XDV0, 085991, Q83VL0
ahdA4	Putative ferredoxin reductase component of dioxygenase	*Sphingomonas yanoikuyae*	A2TC59	[24,32]	1e-40	Q83VI9, A4XDS3, 085962

Organism name and source, accession number and bibliographic reference for each reference protein. BLASTp expected threshold values used and selected sequences for multiple alignments for the probe design are given.

enzyme. Also, to reduce the number of specific probes synthesized on the microarray, the last nucleotide of each probe (generally a degenerate base or an inosine due to degeneracy of the genetic code) is also manually eliminated.

Thus by these strategies, two degenerate probes targeting two different regions are selected per targeted gene (Table 2). Based on these sixteen 23-mer degenerate probes, we finally obtain 8,048 specific probes.

Explorative probe validation

Strain EPA505 is known to utilize PHE and FLA as sole sources of carbon and energy for growth [30]. However, for this strain, the enzymes involved in the catabolism

Table 2 Selected probe information

Targeted Gene	Probe name	Sequence	Number of unique DNA sequences used for the probe design	Number of specific probes	Positions on the reference gene sequence
phnA1a	phnA1a_MD_A	GTITGYAAYTAYCAYGGITGGGT	5	256	294 - 316
	phnA1a_MD_B	CAYGARATHGARGTITGGACITA	4	384	957 - 979
phnA2a	phnA2a_MD_A	GARGAYATHCAYTAYTGGATGCC	2	48	123 - 145
	phnA2a_MD_B	GGICARGTITGGATGGARGAYCC	3	128	261 - 284
ahdA1c	ahdA1c_MD_A	GARTGYGTITAYCAYCARTGGGC	3	128	318 - 340
	ahdA1c_MD_B	GAYGCIGCIGAYAARCARGCITA	2	1024	771 - 793
ahdA2c	ahdA2c_MD_A	GAYGAYMGIYTIGARGARTGGCC	3	1024	081 - 103
	ahdA2c_MD_B	ATHGAYACIATGATGGTIMGICC	3	768	459 - 481
bphB	bphB_MD_A	AAYGTIGGIATHTGGGAYTWYAT	3	768	261 - 283
	bphB_MD_B	AAYBTIAARGGITAYTTYTTYGG	3	384	348 - 370
bphC	bphC_MD_A	CCITAYTTYATGCAYTGYAAYGA	5	128	558 - 580
	bphC_MD_B	TGGYTITGGGARTTYGGITGGGG	4	128	777 - 799
bphA3	bphA3_MD_A	ATHATHGARTGYCCITTYCAYGG	2	576	180 - 202
	bphA3_MD_B	ATHGAIGAYGGITGGGTITGYAT	3	768	279 - 302
ahdA4	ahdA4_MD_A	GCIAAYGTICCIGAYAAYTTYTT	2	1024	159 - 181
	ahdA4_MD_B	CARGARACITAYCARAAYGCIGC	2	512	867 - 889

Total number of specific probes from the probe degenerate sequence and relative positions on the reference gene sequence for each targeted gene are described. Numbers of unique DNA sequences, coding for studied enzymes are also given to highlight that our probes target known genes but also unknown ones. Nomenclature: M: A and C; R: A and G; W: A and T; S: G and C; Y: C and T; H: A, C and T; D: A, G and T; B: G, T and C; I: A, C, G and T.

of PHE and FLA have not been fully characterized. Only gene fragments for the ferredoxin component of dioxygenase (*pbhB* equivalent to *bphA3*) and for the 1,2-dihydroxy-biphenyl-2,3-diol 1,2-dioxygenase (*pbhA* equivalent to *bphC*) are available in public databases [31] for the studied enzymes. This strain is thus an excellent model to validate our approach, as we could work with no prior assumptions using explorative probes to ensure the detection of unidentified genes.

With this aim, growth kinetics experiments with PHE, FLA and a mix of both pollutants as sole carbon and energy source are carried out to evaluate the targeted gene expression. As expected, for the eight genes studied, we have detected positive hybridizations (SNR' > 3) on the DNA microarray using mRNA as targets extracted after 3 h of culture (Table 3). Surprisingly, we do not observe positive signals with the probes targeting one region of the *phnA2a* gene. However, one probe targeting the second region of this gene allow the detection of strong hybridization signals (SNR' = 22.64 ± 3.21) indicating a potentially high level of gene expression induced by PAHs. Additionally, control experiments with glucose as sole carbon and energy source do not give positive hybridizations for most of the targeted genes (Table 3). The SNR' value indicating positive hybridization is close to the threshold reflecting a low gene expression. These results suggest that all the studied genes can be induced in response to the mix of PAH exposure. The same results are obtained for growth kinetics with one PAH (PHE or FLA) as sole carbon and energy source. The same specific probes give the highest SNR' for the eight targeted genes, but with different levels of induction. For example, for the same specific probe (named bphA3_MD_B_0333) targeting the region B of the gene *bphA3* in all PAH-cultures we find: 9.79 ± 1.39 with a mixture of two pollutants, 20.00 ± 5.84 with PHE alone, 7.50 ± 2.03 with FLA alone and no positive signal with glucose. We note that the number of probes giving a positive signal is low for targeted genes (between 1 for *phnA2a* and 5 for *bphB* after 3 h of culture with the mix of PAHs) reflecting variable levels of similarity between targets and probes deduced from variably degenerate regions.

Based on these results, we can also predict the most likely gene sequence of the targets interacting with probes. Among the positive probes, one shows a strong signal (e.g. one targeting *bphA3* with a median SNR' = 36.87 ± 7.83) compared with the others targeting the same region. We hypothesize that the strongest SNR' probe perfectly matched, or is the closest sequence to targeted genes. Using sequences of *bphA3* and *bphC* genes available in databases [EMBL: AF259397 and AF259398], we demonstrate that only two probes among the four have identical sequences with *bphC* and *bphA3* genes. These data do not confirm the efficiency of our approach, and so to validate our first observations, we decide to isolate and characterize these genes and the others by a combination of amplification, cloning and sequencing strategies. Four gene clusters of 4.47 kb, 2.13 kb, 1.20 kb, and 0.32 kb, respectively [EMBL: FM882255, FM882254, FM882253 and FN552592] are thereby obtained. The complete nucleotide sequence of the 4.47 kb contig [EMBL: FM882255] shows six putative non-overlapping open reading frames (ORFs). Among these, four are targeted with our microarray probes. The first encodes a polypeptide 98% similar to a putative biphenyl-2,3-diol 1,2-dioxygenase known to degrade various dihydroxy-PAHs, and named BphC [EMBL: BAC65429]. The second encodes a polypeptide 90% similar to a putative ferredoxin component of dioxygenase, named BphA3 [EMBL: BAC65428], involved in various steps of the process of PAH degradation for the electron transfer from reductase to dioxygenase complex [26]. Interestingly, these two ORFs are highly similar to

Table 3 Results obtained with designed probes for a mixture of phenanthrene and fluoranthene.

Gene name	*phnAla*		*phnA2a*		*ahdAlc*		*ahdA2c*		*bphB*		*bphC*		*bphA3*		*ahdA4*	
Targeted region	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Total number of specific probes	256	384	48	128	128	1024	1024	768	768	384	128	128	576	768	1024	512
Number of specific probes giving a positive signal (SNR' > 3)	1	2	0	1	3	1	2	1	4	1	1	1	3	1	0	0
Highest median SNR' obtained for each targeted region	18.32 ± 3.64	6.62 ± 0.31	X	22.64 ± 3.21	8.61 ± 1.59	9.93 ± 1.32	8.92 ± 1.52	16.26 ± 2.45	5.79 ± 1.73	4.09 ± 0.66	4.47 ± 0.30	4.54 ± 0.81	36.87 ± 7.83	9.79 ± 1.39	X	X
Specific probe for EPA505 gene giving highest median SNR'	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No
For comparison, total number of specific probes giving a positive signal with glucose	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

For each degenerate probe defined targeting two different regions (A and B) of genes (*phnA1a, phnA2a, ahdA1c, ahdA2c, bphB, bphC, bphA3* and *ahdA4*), total number of specific probes stemming from the degenerate sequence, total number of specific probes giving a 'positive' signal (with a SNR' > 3), highest median SNR' visualized for each targeted region of each gene and whether the probe specific to the strain EPA505 gene gives this highest signal median SNR'.

available sequences for strain EPA505 [31], but a comparison with our sequences reveals some mismatches. The two last genes encode two polypeptides respectively 88% and 95% similar to AhdA2c [EMBL: BAC65427] and AhdA1c [EMBL: BAC65426], two components of a terminal oxygenase involved in the monooxygenation of salicylate, a metabolic intermediate of PHE, to catechol [32,33]. Two genes identified on the 2.13 kb contig (FM882254) encode polypeptides of 455 and 175 residues. These polypeptides resemble in length and sequence the alpha (99% sequence identity) and beta (100% sequence identity) subunits [EMBL: CAG17576 and CAG17577] of the ring-hydroxylating dioxygenase (*phnA1a* and *phn2a* respectively) of *Sphingomonas* sp. CHY-1, involved in the conversion of several PAHs into their corresponding dihydrodiols [28,34]. The third contig of 1.20 kb (FM882253) encompasses a single partial ORF encoding a polypeptide displaying 95% similarity with the ferredoxin reductase component of a dioxygenase, named AhdA4 [EMBL: BAC65450] of *Sphingobium* sp. P2 and involved in the electron transfer in association with BphA3 [35]. The last contig of 0.32 kb [EMBL: FN552592] encodes a partial 107 amino acid sequence 97% similar to a 1,2-dihydrodiol-1,2-dihydroxy-dehydrogenase named BphB [EMBL: ABM79802] of *Sphingobium yanoikuyae* B1.

Comparison of these gene sequences with the microarray probes shows that our design strategy is efficient to detect, with no prior sequence assumptions, targeted genes from complete metabolic pathways. As expected, for each gene, different probes give positive signals in agreement with the gene sequence composition. Furthermore, among the thirteen probes (targeting both regions of the eight genes) giving the highest signals, nine probes perfectly match strain EPA505 targeted gene regions (Table 3). Thus the two regions (A and B) selected for *bphA3* and *ahdA2c* genes probe designs allow the specific identification of these genes. For the genes *phnA1a*, *phnA2a*, *ahdA1c*, *bphB* and *bphC*, only one region can be considered specific for the identification of the genes. Finally, for *ahdA4* gene, as no probes give positive signals, we can then hypothesize that *ahdA4* is not expressed or is weakly expressed (under the detection threshold) in our culture conditions. We can also postulate that absence of signal might reflect a low sensitivity of these selected probes targeting *ahdA4*.

To conclude, these results confirm that our design strategy is useful and efficient for the targeted genes studied. These data also show that it is essential to select at least two specific regions for each studied gene that should be experimentally validated to ensure accurate identification. Nevertheless, a majority of selected regions is useful for the design of efficient probes that perfectly hybridize with their targets and show the strongest signal on the microarray.

Gene expression analysis with microarray and quantitative real-time PCR experiments

As described previously, the applied design strategy lets us to detect targeted genes from the studied metabolic pathway without prior assumptions. It is thus of interest to test whether our DNA microarray is able to evaluate mRNA levels semi-quantitatively during biodegradation kinetics with PHE, FLA and a mixture of the two pollutants as sole carbon and energy source. A control experiment with glucose as sole carbon and energy source is also conducted. For these four conditions, total RNAs are extracted from pure cultures of strain EPA505 at different times of the kinetics (0, 3, 6, 10 and 21 h). According to the explorative probe validation conclusions (see previous section), only the most efficient probes targeting each of the eight genes in response to pollutant exposure are considered. In addition, to evaluate the gene expression level, a quantitative reverse transcription PCR approach is also developed for the selected genes during the same times of the kinetics.

Transcript hybridizations obtained with only glucose-amended cultures give no positive probe signals (SNR' > 3) for the different times of the kinetics studied as shown in Additional file 1. Under PHE-growth conditions, specific probes give positive signals (SNR' > 3) after 3 h of growth for all the studied genes (Additional file 1). Detected signals largely decrease at 6 h of culture to reach SNR' values under the set threshold. Same SNR' values, in agreement with absence or low abundance of targeted mRNA, are also obtained after 10 h and 21 h of culture (Additional file 1). With FLA as carbon source, except for *ahdA1c*, *bphC* and *bphB* probes, positive SNR' values are also obtained with specific probes after 3 h of growth. After 6 h of culture with FLA, no positive probe signal (SNR' > 3) is detected, as in glucose-growth conditions (Additional file 1). Surprisingly, a positive signal for the specific probe targeting *bphB* is detected after 6 h of culture (SNR' = 3.43 ± 0.70) with FLA, but not after 3 h of culture. Finally, with a mixture of the two pollutants, high positive signals are detected, except for the *ahdA4* gene, under the SNR' threshold and for *bphC* and *bphB*, just above the SNR' threshold, after 3 h of culture (Additional file 1). After a large decrease in SNR' values after 6 h of culture, positive signals for most of the probes are visualized after 10 h of culture, indicating a new gene expression induction. Finally, at 21 h of culture, the detected signals have the same SNR' values as those obtained with glucose. Gene expression results obtained with microarray assays show an up-regulation of all the studied genes with different mRNA levels according to

PAH exposure (Additional file 1). For *ahdA4*, no positive signals are detected except with PHE after 3 h of culture with a SNR' close to the threshold (SNR' = 3.19 ± 0.40).

At the same time, a quantitative reverse transcription PCR based approach is used to precisely describe the gene expression during the growth kinetics. Results show the same expression profiles as those observed with DNA microarray experiments (Additional file 1). Low mRNA levels are detected during growth on glucose, indicating a very low basal gene expression in the absence of PAH substrates. With PHE or FLA as sole carbon and energy source, a high level of targeted mRNA is detected after 3 h of growth. However, a higher mRNA level is detected with PHE exposure. For these two cultures, after 10 h of culture, gene transcript number decreases to reach mRNA levels close to or below the control copy number detected in glucose-grown cells, as with results visualized with the DNA microarrays. With a mixture of the two pollutants, the same expression profile is detected with the quantitative reverse transcription PCR approach and with the DNA microarrays. High mRNA levels are measured after 3 h of culture, and besides a large decrease after 6 h of culture, another mRNA up-regulation is detected at 10 h of culture for the studied genes. Finally, mRNA levels decrease to reach transcript levels close to growth experiments performed with glucose. In conclusion, similar expression profiles are obtained for *phnA1a*, *phnA2a*, *ahdA1c*, *ahdA2c*, *bphB*, *bphC* and *bphA3* with DNA microarray and quantitative reverse transcription PCR approaches, demonstrating the efficiency of probes designed using Metabolic Design software. Thus DNA microarrays using Metabolic Design can be used to perform semi-quantitative monitoring of gene expression.

Characterization of potential metabolic capacities in a PAH contaminated soil

As we developed explorative probes to detect key genes coding for enzymes involved in PAH degradation, we assess the metabolic capacities of endogenous microbial communities in a polluted ecosystem. Owing to the difficulty in extracting microbial RNA in such environments, we hybridize total extracted microbial DNA from a highly contaminated soil (contamination details in Additional file 2). This ecosystem is selected because it harbors high concentrations of PAHs (2,300 mg/kg of dry soil). Also, PHE and FLA are detected as major contaminants (respectively 430 and 270 mg/kg of dry soil).

Among the 8,048 designed probes targeting the eight genes, 358 give positive signals (SNR' > 3) after hybridization with total DNA (Table 4). For each gene, probe sets show strong signals, but with variable intensities, identifying the most probable target sequence. To evaluate the explorative capacities of our probes, we first focus on the *phnA2a* gene. We compare the signal intensities between mRNA hybridization of strain EPA505 and the DNA extract from the polluted soil (Figure 4). We clearly identify the probe signature for strain EPA505 and a specific probe signature for the polluted soil. Using a BLASTn approach with complete databases (EMBL), 21 positive probe sequences have high similarities (0, 1 or 2 mismatches) with *phnA2a* genes from known PAH degraders (such as *Novosphingobium* sp. H25, *Cycloclasticus* sp. NY93E or *Sphingomonas* sp. CHY-1) (data not shown). We can then hypothesize that other positive probe sequences presenting a slight homology with available *phnA2a* sequences might have targeted *phnA2a* unknown genes, consistent with the explorative purpose of these probes.

The highest SNR' signal is given for a probe targeting *ahdA1c* (42.85 ± 5.83) among 204 other positive probes for this gene. As for *phnA2a* positive probes, several are potentially explorative. Interestingly, specific probe targeting *ahdA1c* gene from strain EPA505 also gives a positive signal (median SNR' = 7.45 ± 0.34). The same positive results are obtained with probes specific to strain EPA505 genes: 3.12 ± 1.00 for *phnA2a*, 4.07 ± 0.27 for *ahdA2c*, 4.33 ± 1.14 for *bphC* and 7.06 ± 1.22 for *bphA3*, suggesting the presence of bacteria closely related to strain EPA505.

Surprisingly, no probe can detect *phnA1a* gene in the polluted soil. We choose to amplify, with a PCR approach, *phnA1a* genes using degenerate primers (data not shown). The PCR products are then cloned, and eight clones are sequenced. Among these eight sequences, seven showing high similarities with *phnA1a* genes are then compared with our probe sequences. This comparison reveals multiple mismatches (data not shown), impeding hybridizations with our probes. This result indicates a marked divergence of this gene family. Our first design focused on *phnA1a* genes related to *Sphingomonas*. For a broader discovery of gene diversity, we will need to design probes that take into account more exhaustively the most complete sequence diversity in databases (international and/or personal).

Discussion

We have developed and validated a new algorithm named Metabolic Design. This software can be used to design efficient explorative probes for functional DNA microarrays. Previously to probe design, users have to extract from public (Swiss-Prot and TrEMBL) or personal databases, protein sequences of interest. Results are then integrated in a user-friendly, intuitive interface. All databases used for the application can be selected by the users and they can also integrate personal data. Such flexibility is generally not available, for example with

Table 4 Results obtained with designed probes with total DNA extracted from the contaminated soil S3

Gene name	*phnAIa*	*phnA2a*	*ahdAIc*	*ahdA2c*		*bphB*	*bphC*	*bphA3*	*ahdA4*
Targeted region	**A**	**B**	**B**	**A**	**B**	**A**	**B**	**B**	**A**
Total number of specific probes	256	128	1024	1024	768	768	128	768	1024
Number of specific probes giving a positive signal (SNR' > 3)	0	37	204	18	1	36	16	44	2
Percentage of probes giving a positive signal (SNR' > 3)	0	28.90	19.92	1.75	0.13	4.68	12.50	5.72	0.19
Highest median SNR' obtained for each targeted region	0 ± 0.00	9.47 ± 0.70	42.85 ± 5.83	7.05 ± 1.37	4.29 ± 1.71	6.33 ± 2.05	4.43 ± 1.31	8.84 ± 2.15	3.48 ± 0.98

For each degenerate probe defined targeting one particular region (A or B) of genes (*phnA1a, phnA2a, ahdA1c, ahdA2c, bphB, bphC, bphA3* and *ahdA4*), total number of specific probes stemming from the degenerate sequence, total number of specific probes giving a 'positive' signal (with a SNR' > 3), probe percentage giving a 'positive' signal and highest signal median SNR' visualized for each targeted region of each gene.

current metabolic reconstruction tools, such as the 'Pathway Tools Software', initially developed for the EcoCyc project [36], or KEGGanim [37]. These are generally based on static databases and predefined metabolic pathways (such as KEGG [38], MetaCyc [39] or BRENDA [40]).

In order to bypass the faulty annotations found in automatically filled databases, and to allow the exhaustive exploitation of all the currently available protein sequences, the mining step is performed using similarity search. However, such approach presents another major drawback. Indeed, in some cases, not all proteins with a similar function have similar primary structures. Thus a future development of Metabolic Design will be the replacement of the BLASTp step by a Pattern Hit Initiated BLAST (or PHI-BLAST) step coupled with PRODOM data (defined as a comprehensive set of protein domain families automatically generated from the UniProt Knowledge Database) [41]. PHI-BLAST analysis is useful for identifying the distant members of a protein family, whose relationship is not recognizable by straight sequence comparison, but only by patterns contained in sequences (such as catalytic sites or substrate recognition sites). We also intend to integrate a new module for high-throughput ortholog prediction (using for example Ortho-MCL or Ortholuge) to improve homologous protein selection for complex and divergent protein families [42,43].

The ultimate aim of Metabolic Design is to define explorative probes and estimate their specificity *in silico*. Specific probes deduced from defined degenerate probes thus allow the targeting not only of known gene sequences but also of new ones that encode the same protein sequences. These explorative features are not offered by other tools such as OligoArray 2.0, YODA or HPD [17]. In addition, Metabolic Design takes into consideration both *ex situ* and *in situ* DNA microarray synthesis. The inosine composition is taken into account in the total degeneracy, as an *ex situ* microarray can hold inosine nucleotide probes, and/or degenerate probes in one spot, reducing probe degeneracy.

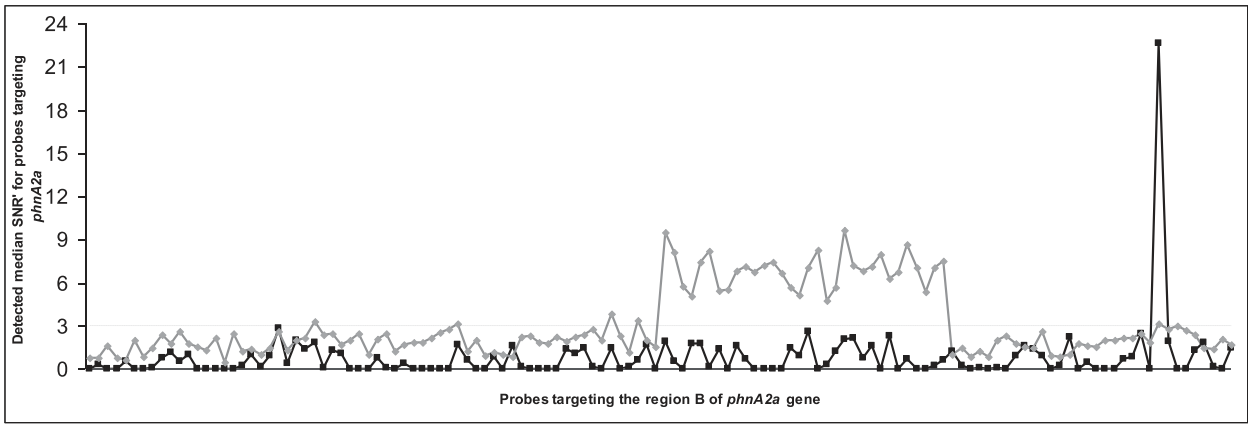


Figure 4 Median SNR' for the contaminated soil with 128 specific probes targeting the *phnA2a* gene. This graphic represents the detected median SNR' for each specific probe (ordered by sequence) derived from the degenerate defined probe **phnA2a_MD_B** targeting one particular region of *phnA2a* gene. Black squares: signals obtained with the model strain EPA505 with a mix of both pollutants (the highest signal is given by the specific probe targeting the strain EPA505 specific gene). Gray diamonds: signals obtained with total DNA extracted from the soil S3 (clearly showing a particular probe signature). The dotted line represents the defined threshold for SNR' values.

Probe specificity is then evaluated *in silico* using a proprietary database, giving us a close glimpse of potential cross-hybridizations found in complex environments. In addition, in Metabolic Design, this database can be modified to consider complete DNA data, or only fragmented data (for example, only one genome). Estimation and validation of potential cross-hybridizations are performed by a BLASTn analysis. However, one possible improvement would be to take into account optimized BLASTn parameters recently described as allowing a more efficient detection of potential cross-hybridizations [44].

Another update of Metabolic Design will add thermodynamic calculations to improve probe selection, although these parameters are not fully described at present [21,45]. Also, it will be essential to take into account probe sensitivity due to sequence nature [46]. In view of these difficulties in precisely predicting probe behavior during DNA microarray hybridizations, we suggest that users first validate the quality of the DNA microarrays (probe specificity and sensitivity), with a simple biological model as we did in this study.

Based on Metabolic Design defined probes, targeting eight genes coding for enzymes involved in the degradation of various PAHs by strain EPA505, we demonstrate that our design strategy is useful for most of the determined probes. Furthermore, these results highlight the capacity of our probes for semi-quantitative monitoring of gene expression or gene detection, confirming the quantitative capability of our microarrays for environmental applications [14]. Finally, we demonstrate the explorative ability of our probes, studying a complex environment. Indeed, most classical functional microarrays (such as GeoChip) using specific probes will monitor only known sequences and cannot appraise the complete microbial gene diversity of complex environments [13,14,47,48]. Additionally, considering the high complexity of environmental samples, it will be interesting to improve again probes specificity and sensitivity, using for example the 'GoArrays' strategy [29].

To allow the identification of complete sequences of targeted genes, a further application of these explorative DNA microarrays will be the capture of 'unknown' sequences for further next-generation sequencing [49,50]. Some new techniques have been reported for performing selective capture of sequence fragments from complex mixtures based on hybridization to DNA microarrays. Combining our explorative DNA microarrays and next-generation sequencing will, for example, bypass a critical bottleneck in microbial ecology, namely the difficulty of specifically exploring some biochemical pathways or specific biomarkers without the need to sequence the complete metagenome or PCR products (not reflecting reality due to PCR artifacts). Most often in complex environments even with high throughput sequencing, we obtain only a partial view of the extremely broad microbial diversity. In addition, using mRNA or large DNA fragments as targets can allow all the genes included in a transcriptional unit to be captured. So, in prokaryotes, like genes involved in the same biological process are generally associated in the same transcriptional unit, this capture would allow to assign of new gene functions.

Conclusions

This study evaluates the efficiency of a new probe design software tool, Metabolic Design, dedicated to DNA functional microarrays. This software, which can be used to study any group of genes, was successfully applied to define probes able to detect with high specificity and sensitivity genes encoding enzymes involved in PAH degradation. In addition, DNA microarray experiments performed on soil polluted by organic pollutants, without prior sequence assumptions, demonstrate explorative abilities of our probes. So, probe design performed with Metabolic Design ensures to precisely monitor metabolic regulations during various processes in complex environments.

Methods

Software implementation

The Metabolic Design application can be obtained on request *via* FTP and runs only on MS-WINDOWS (32-bit) platforms. The Java runtime environment (JRE) Version 1.4 or higher, Perl Version 1.5 or higher and an SQL database such as Oracle 9i must be installed. Latest Swiss-Prot and TrEMBL database versions have also to be downloaded for local installation of data from ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/complete. Metabolic Design is a standalone multilayered tool comprising a relational database, a data object layer and a graphical user interface (GUI). For the data mining step, curated information associated with enzymes (systematic name and source organism) come from the international protein database Swiss-Prot/TrEMBL and/or personal data. Swiss-Prot/TrEMBL data are parsed using PERL scripts allowing extraction of files including sequences from prokaryotes and fungi. To evaluate potential cross-hybridization of candidate probes, microbial related sequences from the EMBL database, which include environmental samples (ENV), fungi (FUN) and prokaryote (PRO) taxonomic divisions, are selected. DNA sequences corresponding to CDSs with their respective putative 5' and 3' UTR flanking regions (arbitrarily set at 100 nt each) are then extracted and formatted to perform the cross-hybridization checking step by a tBLASTn approach ('cross-hybridization database'). For particular BLAST steps, some parameters

are defined for the Metabolic Design program: tBLASTn for the specificity step (*-e 10000000 -w 2 -b 5000 -v 5000*), BLASTx for the comparison with the reference protein sequence (*-e* $1e^{-10}$ *-F F*), BLASTn for Kane's criteria evaluation (*-e 10 -w 7 -F F -q -1*), and BLASTn for cross-hybridization clustering (*-e* $10e^{-10}$ *-w 7*).

A JAVA classes package is developed to implement the data object layer and GUI. Object persistence is guaranteed at both text file and SQL levels.

Chemicals

PHE, FLA, Tween 80, HPLC grade solvents and acetone are purchased from Sigma-Aldrich (Saint-Quentin-Fallavier, France). For degradation experiments, stock solutions of each PAH (PHE and FLA) are prepared in acetone at a final concentration of 2 g/L and sterilized as described above. A mixture of PHE and FLA (1 g/L each) is prepared in the same way.

DNA extraction from soil

Total DNA is extracted from 5 g of contaminated soil (S3) following the protocol described by Zhou [51]. Three extracts are made and pooled to minimize potential biases. DNA quality is checked on a 0.8% agarose gel.

Bacteria, growth conditions and kinetic experiments

Strain EPA505 (DSM7526) is purchased from DSMZ (Braunschweig, Germany). Cells are first grown overnight at 37°C on a shaker table (150 rpm) in 70 mL of Luria-Bertani medium (LB) containing streptomycin (100 mg/L) to produce biomass. They are then centrifuged at room temperature for 2 min at 5,000 *g* and transferred to a sterile minimum mineral medium 457 containing, for 1 L of distilled water, 2,440 mg of Na_2HPO_4, 1,520 mg of KH_2PO_4, 500 mg of $(NH_4)_2SO_4$, 200 mg of $MgSO_4$ - $7H_2O$, 50 mg of $CaCl_2$ - $2H_2O$, 200 mg of Tween 80 and 10 mL of SL-4 solution as described by DSMZ [30,33]. Cultures are prepared as follows: 2 mL of the PAH stock solutions is evaporated in sterile 250 mL conical flasks, 100 mL of sterile medium 457 is added, and the flasks are inoculated with 7.5 $\times 10^7$ cells prepared as above. These cultures are incubated at 28°C on a shaker table (150 rpm) for 27 h and bacterial growth is monitored spectrophotometrically at 620 nm using an Ultraspec 2000 spectrophotometer (Pharmacia Biotech AB, Uppsala, Sweden). A culture is also grown with glucose (15 g/L) as sole carbon source and energy to define the basal expression of genes implicated in PAH degradation.

RNA extraction from strain EPA505

Total RNA from a pure culture of strain EPA505 is extracted at different times of the PAH degradation kinetics (0, 3, 6, 10, and 21 h) with the RNeasy Mini kit (Qiagen GmbH), and treated with 1.5 U of DNase I (Invitrogen) to eliminate DNA contamination. RNA sample concentration and purity are then estimated using a Nanodrop spectrophotometer (Nanodrop).

Microarray experiments

Samples of 15 μL of strain EPA505 total RNA of four PAH degradation kinetics data points (0, 3, 6, 10 and 21 h) are enriched using the MICROB*Express*™ Bacterial mRNA Enrichment Kit (Ambion) as recommended by the suppliers. Each enriched mRNA is then amplified using the MessageAmp™ II-Bacteria RNA Amplification Kit (Ambion) with a modified protocol for the *in vitro* transcription step. Briefly, the purified double-stranded template (~14 μL) is transcribed *in vitro* with 12 μL of ATP, CTP and GTP mix (25 mM each) (Ambion), 3 μL of UTP (75 mM) (Ambion), 3 μL of amino-allyl-UTP (50 mM) (Ambion), 4 μL of 10 × reaction buffer (Ambion) and 4 μL of T7 enzyme mix (Ambion) at 37° C for a 14 h incubation period. Finally, the aRNA is purified using the MessageAmp™ II-Bacteria RNA Amplification Kit (Ambion) following the manufacturer's instructions.

In the next step, 10 μg of purified aRNA for each sample are vacuum-dried and labeled using the Amersham CyDye™Post-Labeling Reactive Dye Packs (GE Healthcare, Little Chalfont, United Kingdom) with Cyanine3 or Cyanine5 dyes as recommended by the supplier. Briefly, the aRNA pellet is resuspended in 20 μL of 0.1 M bicarbonate buffer (pH 8.7) and incubated for 90 min with 40 nM of dye compound (coupling the dye to amino-allyl-UTP) dissolved in 20 μL of DMSO (dimethyl sulfoxide) in the dark at room temperature. Excess dye is quenched by adding 15 μl of 4 M hydroxylamine solution incubated for 15 min in the dark at room temperature. The labeled aRNA is then purified with NucleoSpin RNA Clean-Up kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's instructions. After each step (total RNA enrichment, RNA amplification and aRNA labeling), the quantity and integrity of RNA are estimated using the RNA 6000 Nano kit (Agilent Technologies), the Agilent 2100 Bioanalyzer (Agilent Technologies) and the Nanodrop spectrophotometer (Nanodrop) as recommended by protocols.

Total DNA is amplified and labeled using the BioPrime® Total Genomic Labeling System (Invitrogen) following the manufacturer's instructions. The quantity and quality of labeling are estimated using a Nanodrop spectrophotometer (Nanodrop) as recommended by protocols.

NimbleGen custom arrays of 8,048 probes are used (Roche NimbleGen, Madison, USA). All the probes

are randomly distributed across the array to minimize spatial effects as far as possible during the hybridization step. The microarray also contains thousands of random probes (randomly defined length and sequence) which can serve to measure technical background noise. For each hybridization experiment, 3.33 μg of labeled RNA for kinetic experiments (one sample in Cy3 and another in Cy5) or 12 μg of labeled DNA (in Alexa Fluor® 5) for soil are mixed, vacuum-dried and resuspended in 5.6 μL of water. The hybridization mix (Roche NimbleGen) is then made according to the manufacturer's protocols. The arrays are hybridized on a 4-bay NimbleGen Hybridization System (Roche NimbleGen) at 42°C for 72 h. The arrays are washed with NimbleGen wash buffers I, II and III according to vendors' protocols and scanned using a Scanner Innoscan 900AL (Innopsys, Carbonne, France) at 2 μm resolution. Individual array images are acquired independently, adjusting the PMT gain for each image as recommended using Mapix® software (Innopsys).

For each array image, raw expression data are extracted using the NimbleScan software v2.1. (Roche NimbleGen) and feature intensities are exported as .pair files. The background noise is then determined using random probes present on the microarrays (8,863 probes in our experiment) with the method described in the Additional file 3. This background noise is defined by two components: the background median intensity (Bposition) and its dispersion (Bdispersion). Finally, a modified signal-to-noise ratio termed SNR' and based on the formula of Verdik [52] is calculated as follows in order to reduce-centralize our data: SNR' = (probe signal intensity - Bposition)/Bdispersion (see Supplementary Data S1).

However, spatial effect across the array surface is a predominant within-slide experimental artifact that needs to be eliminated before any other normalization procedure [53]. Accordingly, for all array images obtained in this work, the surface is segmented into 16 sub-squares according to probe position (*X*, *Y*) indicated in the pair report. A Perl script is developed to calculate local background noise in all sub-squares and the median SNR' retrieved from the three replicates of each probe. Finally, another Perl script is implemented to summarize each replicate probe treated and determine the median value of the three replicates. 'Positive' hybridization is considered significant for probes with SNR' > 3 (value to avoid all false positives) [54]. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE21402: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=rjmhhgoaqyqqinw&acc=GSE21402

DNA extraction, PCR amplification and cloning

Total DNA from a pure culture of strain EPA505 is extracted by heat shocking cells [55]. All PCR reactions are carried out in 50 μL of mixtures containing 20 ng of the previous strain EPA505 DNA extract, 0.5 U of GoTaq DNA polymerase (Promega Corp., Madison, USA), Promega buffer, 1.25 mM $MgCl_2$, 1 μM of each primer depending on the targeted gene (Additional files 4 and 5) and 0.5 mM of each deoxynucleotide. The reactions are performed in a iCycler thermal cycler (Biorad Laboratories, Marnes-la-Coquette, France) using an initial denaturation step consisting of 95°C for 5 min, followed by 35 cycles of 95°C for 1 min, an annealing step with temperature and time depending on primers (Additional Files 4 and 5) and an elongation step of 72° C for 1 min. A final treatment of 72°C for 7 min is then applied. The size and purity of PCR products are checked on 1.2% gel agarose. The PCR products are purified with a Qiaquick Gel Elution kit (Qiagen GmbH, Hilden, Germany), and then ligated into the pCRII-TOPO® vector supplied with the TA cloning kit (Invitrogen Corp., Merelbeke, Belgium) and transformed into *E. coli* One Shot® TOP10 cells (Invitrogen Corp.) following the manufacturer's instructions. White colonies are picked and grown in LB medium supplemented with kanamycin at 50 μg/ml final concentration. Plasmid template DNA is prepared by the alkaline lysis method [55]. The clone inserts are sequenced by the MWG Biotech Company (Ebersberg, Germany) using both SP6 and T7 sequencing primers. Sequence treatment and joining are performed using the pregap4 and the gap4 tools of the Staden Package Program [56]. The gene sequences are then compared with Swiss-Prot and TrEMBL databases using the BlastX program [57].

Real-time PCR experiments

The reverse transcription reactions are carried out at 42° C for 2 h with 50 ng of total RNA using *bphC, bphA3, ahdA2c* and *ahdA1c* (0.625 μM of each primer) mix primers, and *ahdA4, phnA1a, phnA2a and bphB* mix primers respectively (see Additional file 6) in order to minimize manipulation biases. These reactions are carried out in a final volume of 20 μL with 100 U of SuperScriptIII reverse transcriptase (Invitrogen Corp.), 1 U of RNasin+ Inhibitor (Promega Corp.), 0.25 mM of dNTPs mix (Invitrogen Corp.), 0.1 M DTT (Invitrogen Corp.) and Invitrogen buffer, according to the manufacturer's instructions. Reverse transcription reactions are performed in triplicate. cDNA is then diluted ten-fold for quantitative real-time PCR assays. Reactions are carried out with the MESA Green qPCR for SYBR assays kit (Eurogentec) according to the manufacturer's instructions. All amplifications are carried out in a final volume of 20 μL containing 5 μL of sample described above or 5

μL of standard cDNA (from 4.37 × 10^7 copies/μL to 4.37 copies/μL, covering 8 log of dynamic range for each gene), 10 μL of 2× MESA Green qPCR for SYBR assays mixture and the corresponding primers sets described in the Additional file 6 at 0.2 μM final concentration each. The reverse transcription product in kinetic experiment samples is quantified twice. As every reverse transcription experiment is done in triplicate, six measurements are obtained for each sample. Each point on the standard curve (corresponding to serially diluted cDNA) is quantified in triplicate. PCR is carried out in the Mastercycler Realplex (Eppendorf, Le Percq, France) for 1 cycle at 95°C for 5 min followed by 40 cycles consisting of 95°C for 15 s (denaturation step) and 68°C for 45 s (annealing and elongation steps). At the end of the real-time PCR, a melting curve is defined by measurement of SYBR Green signal intensities for 20 min from 68°C to 95°C. Size of the amplified products is checked on 2.5% agarose gel. Data analysis is carried out with *realplex* software (version 1.5; Eppendorf).

Nucleotide sequence accession numbers

The nucleotide sequences reported in this study have been deposited in the database under accession numbers: [EMBL: FM882255] (encompassing *bphC, bphA3, ahdA2c* and *ahdA1c* gene sequences), [EMBL: FM882254] (encompassing *phnA1a* and *phnA2a* gene sequences), [EMBL: FM882253] (encompassing *ahdA4* gene sequences) and [EMBL: FN552592] (encompassing *bphB* gene sequences).

List of abbreviations

GUI: graphical user interface; IUPAC: international union of pure and applied chemistry; PAH: polycyclic aromatic hydrocarbon; PHE: phenanthrene; FLA: fluoranthene; SNR': signal to noise ratio; ORF: open reading frame; CDS: coding DNA sequence; RT-PCR: reverse transcription- polymerase chain reaction; FTP: file transfer protocol; JRE: java runtime environment; SQL: structured query language; UTR: untranslated region, aRNA: antisense RNA; DMSO: dimethyl sulfoxide; cDNA: complementary DNA.

Availability and Requirements

Project name: Metabolic Design

Project homepage: ftp://195.221.123.90/

Operating system: Windows (32-bit) only

Programming language: Java and Perl

Others: The Java runtime environment (JRE) Version 1.4 or higher, Perl Version 1.5 or higher and an SQL database such as Oracle 9i must be installed.

License: Free for non-commercial use. Source code available upon request.

Additional material

Additional file 1: SNR' profiles detected with microarray experiments and transcript numbers profiles detected with quantitative RT-PCR assays. SNR' profiles detected with microarray experiments (LEFT), and transcript copy number detected per ng of total RNA with quantitative RT-PCR assays (RIGHT) for eight genes: (A) *phnA1a*; (B) *phnA2a*; (C) *ahdA1c*; (D) *ahdA2c*; (E) *bphB*; (F) *bphC*; (G) *bphA3*; and ahdA4 (H) during PAH biodegradation at different times with strain EPA505. PHE: grey squares, FLA: triangles, PHE + FLA: circles, glucose: open diamond. Error bars indicate the standard deviation of measures.

Additional file 2: PAH composition detected in the contaminated soil S3. These data are proprietary data given by BioBasic Environnement and give the quantity of detected PAHs in mg/kg of dry soil in the contaminated soil studied.

Additional file 3: Background noise calculation description. Background noise is determined according to 'RANDOM probes response' of Nimblegen microarrays. Our method takes into account the background noise which is characterized by two components: its position and its dispersion.

Additional file 4: Identification of four catabolic genes clusters from the model strain EPA505. Physical maps of four clusters (A, B, C and D) of catabolic genes involved in PAHs biodegradation from strain EPA505. Size of genes and intergenic spaces is indicated as well as position of primers used for PCR amplifications.

Additional file 5: Primer sets used for detecting catabolic genes involved in PAHs degradation and to generate the gene DNA matrix. The DNA matrix is used to build the standard curve for quantitative real-time PCR assays in strain EPA505. *: *xylX* and *nahD* are used to characterize complete sequences of *bphC* and *ahdA1c*. Nomenclature: **M**: A and C; **R**: A and G; **W**: A and T; **S**: G and C; **Y**: C and T; **K**: G and T; **V**: A, G and C; **H**: A, C and T; **D**: A, G and T; **B**: G, T and C; **I**: A, C, G and T.

Additional file 6: Primers used for reverse transcription and quantitative real-time PCR assays. List of primers used for reverse transcription and subsequent quantitative real-time PCR assays. Amplification sizes are also given for each targeted gene.

Acknowledgements

We thank bioinformatics undergraduate students (IUT of Aurillac, Université d'Auvergne) Yann Keriou and Xavier Brotel for PERL script developments and Mathieu Roudel for FTP development, Mélanie Mitchell, an undergraduate student in biology (IUT of Clermont, Université d'Auvergne) for her help in quantitative reverse transcription PCR experimentation, and Brigitte Chebance for technical assistance. ST was supported by a doctoral grant from 'Ministère de l'Enseignement Supérieur et de la Recherche Scientifique'.

Author details

[1]Clermont Université, Université d'Auvergne, Laboratoire: Microorganismes Génome et Environnement, BP 10448, F-63000 CLERMONT-FERRAND, France. [2]CNRS, UMR 6023, Laboratoire: Microorganismes Génome et Environnement, F-63173 AUBIERE, France. [3]Clermont Université, Université Blaise Pascal, Laboratoire: Microorganismes Génome et Environnement, BP 10448, F-63000 CLERMONT-FERRAND, France. [4]Biobasic Environnement, Biopôle Clermont-Limagne, 63360 Saint-Beauzire, France.

Authors' contributions

ST, EP, AM and OG carried out the experimentations and have participated to analysis and interpretation of data. FG, EP, ST and OG have made script developments, updates and optimizations of Metabolic Design. ST, EDB, FG, EP and OG performed *in silico* analysis. ST, EP, OG and PP planned the study and wrote the manuscript. JT, EP, OG, DB, CBP and PP have given final approval of the version to be published. All authors read and approved the final manuscript.

Received: 18 May 2010 Accepted: 23 September 2010
Published: 23 September 2010

References

1. Vieites JM, Guazzaroni M-E, Beloqui A, Golyshin PN, Ferrer M: **Metagenomics approaches in systems microbiology.** *FEMS Microbiol Rev* 2009, **33(1)**:236-255.
2. Schloss PD, Handelsman J: **Toward a Census of Bacteria in Soil.** *PLoS Comput Biol* 2006, **2(7)**:e92.
3. Torsvik V, Øvreås L: **Microbial diversity and function in soil: from genes to ecosystems.** *Curr Opin Microbiol* 2002, **5(3)**:240-245.
4. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, *et al*: **Functional metagenomic profiling of nine biomes.** *Nature* 2008, **452(7187)**:629-632.
5. Urisman A, Fischer K, Chiu C, Kistler A, Beck S, Wang D, DeRisi J: **E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns.** *Genome Biol* 2005, **6(9)**:R78.
6. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: **Microarray-based detection and genotyping of viral pathogens.** *Proc Natl Acad Sci USA* 2002, **99(24)**:15687-15692.
7. Warnecke F, Hess M: **A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts.** *J Biotechnol* 2009, **142(1)**:91-95.
8. Artamonova I, Frishman G, Frishman D: **Applying negative rule mining to improve genome annotation.** *BMC Bioinformatics* 2007, **8(1)**:261.
9. Galperin MY, Koonin EV: **'Conserved hypothetical' proteins: prioritization of targets for experimental study.** *Nucleic Acids Res* 2004, **32(18)**:5452-5463.
10. Singh BK: **Exploring microbial diversity for biotechnology: the way forward.** *Trends Biotechnol* 2010, **28(3)**:111-116.
11. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29(1)**:11-16.
12. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28(1)**:45-48.
13. He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, *et al*: **GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes.** *ISME J* 2007, **1(1)**:67-77.
14. Rhee S-K, Liu X, Wu L, Chong SC, Wan X, Zhou J: **Detection of Genes Involved in Biodegradation and Biotransformation in Microbial Communities by Using 50-Mer Oligonucleotide Microarrays.** *Appl Environ Microbiol* 2004, **70(7)**:4303-4317.
15. Wu L, Liu X, Schadt CW, Zhou J: **Microarray-Based Analysis of Subnanogram Quantities of Microbial Community DNAs by Using Whole-Community Genome Amplification.** *Appl Environ Microbiol* 2006, **72(7)**:4931-4941.
16. Gao H, Yang ZK, Gentry TJ, Wu L, Schadt CW, Zhou J: **Microarray-Based Analysis of Microbial Community RNAs by Whole-Community RNA Amplification.** *Appl Environ Microbiol* 2007, **73(2)**:563-571.
17. Lemoine S, Combes F, Le Crom S: **An evaluation of custom microarray applications: the oligonucleotide design challenge.** *Nucleic Acids Res* 2009, **37(6)**:1726-1739.
18. Jabado OJ, Palacios G, Kapoor V, Hui J, Renwick N, Zhai J, Briese T, Lipkin WI: **Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments.** *Nucleic Acids Res* 2006, **34(22)**:6605-6611.
19. Wang X, Seed B: **Selection of oligonucleotide probes for protein coding sequences.** *Bioinformatics* 2003, **19(7)**:796-802.
20. Nordberg EK: **YODA: selecting signature oligonucleotides.** *Bioinformatics* 2005, **21(8)**:1365-1370.
21. Pozhitkov AE, Tautz D, Noble PA: **Oligonucleotide microarrays: widely applied poorly understood.** *Brief Funct Genomics* 2007, **6(2)**:141-148.
22. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays.** *Nucleic Acids Res* 2000, **28(22)**:4552-4557.
23. Missaoui M, Hill D, Peyret P: **A comparison of algorithms for a complete backtranslation of oligopeptides.** *Int J Comput Biol Drug Des* 2008, **1(1)**:26-38.
24. Keck A, Conradt D, Mahler A, Stolz A, Mattes R, Klein J: **Identification and functional analysis of the genes for naphthalenesulfonate catabolism by Sphingomonas xenophaga BN6.** *Microbiology* 2006, **152(7)**:1929-1940.
25. Ní Chadhain S, Moritz E, Kim E, Zylstra G: **Identification, cloning, and characterization of a multicomponent biphenyl dioxygenase from Sphingobium yanoikuyae B1.** *J Ind Microbiol Biotechnol* 2007, **34(9)**:605-613.
26. Pinyakong O, Habe H, Yoshida T, Nojiri H, Omori T: **Identification of three novel salicylate 1-hydroxylases involved in the phenanthrene degradation of Sphingobium sp. strain P2.** *Biochem Biophys Res Commun* 2003, **301(2)**:350-357.
27. Romine MF, Stillwell LC, Wong K-K, Thurston SJ, Sisk EC, Sensen C, Gaasterland T, Fredrickson JK, Saffer JD: **Complete Sequence of a 184-Kilobase Catabolic Plasmid from Sphingomonas aromaticivorans F199.** *J Bacteriol* 1999, **181(5)**:1585-1602.
28. Demaneche S, Meyer C, Micoud J, Louwagie M, Willison JC, Jouanneau Y: **Identification and Functional Analysis of Two Aromatic-Ring-Hydroxylating Dioxygenases from a Sphingomonas Strain That Degrades Various Polycyclic Aromatic Hydrocarbons.** *Appl Environ Microbiol* 2004, **70(11)**:6714-6725.
29. Rimour S, Hill D, Militon C, Peyret P: **GoArrays: highly dynamic and efficient microarray probe design.** *Bioinformatics* 2005, **21(7)**:1094-1103.
30. Mueller J, Chapman P, Blattmann B, Pritchard P: **Isolation and characterization of a fluoranthene-utilizing strain of *Pseudomonas paucimobilis*.** *Appl Environ Microbiol* 1990, **56(4)**:1079-1086.
31. Story SP, Parker SH, Kline JD, Tzeng TR, Mueller JG, Kline EL: **Identification of four structural genes and two putative promoters necessary for utilization of naphthalene, phenanthrene, fluoranthene by Sphingomonas paucimobilis var. EPA505.** *Gene* 2000, **260(1-2)**:155-169.
32. Cho O, Choi KY, Zylstra GJ, Kim YS, Kim SK, Lee JH, Sohn HY, Kwon GS, Kim YM, Kim E: **Catabolic role of a three-component salicylate oxygenase from Sphingomonas yanoikuyae B1 in polycyclic aromatic hydrocarbon degradation.** *Biochem Biophys Res Commun* 2005, **327(3)**:656-662.
33. Pinyakong O, Habe H, Supaka N, Pinpanichkarn P, Juntongjin K, Yoshida T, Furihata K, Nojiri H, Yamane H, Omori T: **Identification of novel metabolites in the degradation of phenanthrene by *Sphingomonas* sp. strain P2.** *FEMS Microbiol Lett* 2000, **191(1)**:115-121.
34. Jakoncic J, Jouanneau Y, Meyer C, Stojanoff V: **The crystal structure of the ring-hydroxylating dioxygenase from *Sphingomonas* CHY-1.** *FEBS J* 2007, **274(10)**:2470-2481.
35. Pinyakong O, Habe H, Omori T: **The unique aromatic catabolic genes in sphingomonads degrading polycyclic aromatic hydrocarbons (PAHs).** *J Gen Appl Microbiol* 2003, **49(1)**:1-19.
36. Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, *et al*: **EcoCyc: A comprehensive view of Escherichia coli biology.** *Nucleic Acids Res* 2009, **37(suppl_1)**:D464-470.
37. Adler P, Reimand J, Janes J, Kolde R, Peterson H, Vilo J: **KEGGanim: pathway animations for high-throughput data.** *Bioinformatics* 2008, **24(4)**:588-590.
38. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, *et al*: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36(suppl_1)**:D480-484.
39. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2010, **38(suppl_1)**:D473-479.
40. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: **BRENDA, the enzyme database: updates and major new developments.** *Nucleic Acids Res* 2004, **32(suppl_1)**:D431-433.
41. Corpet F, Gouzy J, Kahn D: **The ProDom database of protein domain families.** *Nucleic Acids Res* 1998, **26(1)**:323-326.
42. Chen F, Mackey AJ, Stoeckert CJ, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34(suppl_1)**:D363-368.
43. Fulton D, Li Y, Laird M, Horsman B, Roche F, Brinkman F: **Improving the specificity of high-throughput ortholog prediction.** *BMC Bioinformatics* 2006, **7(1)**:270.
44. Eklund AC, Friis P, Wernersson R, Szallasi Z: **Optimization of the BLASTN substitution matrix for prediction of non-specific DNA microarray hybridization.** *Nucleic Acids Research* 2009, **38(4)**:e27.

45. Mueckstein U, Leparc G, Posekany A, Hofacker I, Kreil D: **Hybridization thermodynamics of NimbleGen Microarrays.** *BMC Bioinformatics* 2010, **11(1)**:35.
46. Royce TE, Rozowsky JS, Gerstein MB: **Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification.** *Nucleic Acids Res* 2007, **35(15)**:e99.
47. Liang Y, Li G, Van Nostrand JD, He Z, Wu L, Deng Y, Zhang X, Zhou J: **Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field.** *FEMS Microbiol Ecol* 2009, **70(2)**:324-333.
48. Liang Y, Nostrand JDV, Wang J, Zhang X, Zhou J, Li G: **Microarray-based functional gene analysis of soil microbial communities during ozonation and biodegradation of crude oil.** *Chemosphere* 2009, **75(2)**:193-199.
49. Bau S, Schracke N, Kränzle M, Wu H, Stähler P, Hoheisel J, Beier M, Summerer D: **Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays.** *Anal Bioanal Chem* 2009, **393(1)**:171-175.
50. Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stähler CF, Chee MS, Stähler PF, Beier M: **Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing.** *Genome Res* 2009, **19(9)**:1616-1621.
51. Zhou J, Bruns MA, Tiedje JM: **DNA recovery from soils of diverse composition.** *Appl Environ Microbiol* 1996, **62(2)**:316-322.
52. Verdick D, Handran S, Pickett S: **Key considerations for accurate microarray scanning and image analysis.** In *DNA array image analysis: nuts and bolts.* Edited by: Kamberova G. LLC DP: Salem, MA; 2002:83-98.
53. Wang X, He H, Li L, Chen R, Deng XW, Li S: **NMPP: a user-customized NimbleGen microarray data processing pipeline.** *Bioinformatics* 2006, **22(23)**:2955-2957.
54. He Z, Zhou J: **Empirical Evaluation of a New Method for Calculating Signal-to-Noise Ratio for Microarray Data Analysis.** *Appl Environ Microbiol* 2008, **74(10)**:2957-2966.
55. Sambrook J, Fritsch E, Maniatis T: **Molecular cloning: A Laboratory Manual - Third Edition.** Cold Spring Laboratory Harbor Press, U.S.A 2001.
56. Staden R: **The staden sequence analysis package.** *Mol Biotechnol* 1996, **5(3)**:233-241.
57. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3)**:403-410.

doi:10.1186/1471-2105-11-478
Cite this article as: Terrat *et al.*: Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics* 2010 **11**:478.